Journal of Computing in Civil Engineering

NLP-based approach to classifying heterogeneous terms for unambiguous exchange of roadway data

--Manuscript Draft--

Manuscript Number:	CPENG-1952R1
Full Title:	NLP-based approach to classifying heterogeneous terms for unambiguous exchange of roadway data
Manuscript Region of Origin:	UNITED STATES
Article Type:	Technical Paper
Funding Information:	
Abstract:	The inconsistency of data terminology due to the fragmented nature of the highway industry has imposed big challenges on integrating digital data from distinct sources. The issue of semantic heterogeneity may lead to the lack of common understanding of the same data between the sender and receiver. Explicit semantic relations among terms in digital dictionaries, such as ontologies can enable the meaning of a roadway concept name to be transparent and unambiguously understood by computer systems. However, due to the lack of an effective automated method, current practices of identifying these relations hardly rely on a manual process of knowledge acquisition from domain experts or text documents which is laborious and time-consuming. This paper presents a novel methodology that leverages recent advances in Natural Language Processing (NLP) techniques to extract English-American roadway terms used in different government agencies and their semantic relations from roadway design manuals and specifications. The proposed method includes the following three stages: (1) implementing NLP techniques to detect commonly used technical terms from the highway corpus; (2) utilizing machine learning to learn the semantic similarity among roadway terms using their context data in the corpus; and (3) developing a classification algorithm to identify semantic relation types among technical terms. The key merit in this technique is the automated identification of semantic relations among heterogeneous roadway terms from design guidebooks without reliance on other existing hand-coded semantic resources. The proposed methodology was evaluated by conducting an experiment comparing the automatically-identified synonyms by the proposed system with a human-constructed golden standard dataset obtained from Wikipedia. The result shows that the proposed model achieves a precision of over 80 percent.
Corresponding Author:	H. David Jeong, Ph.D. Iowa State University Stillwater, OK UNITED STATES
Corresponding Author E-Mail:	djeong@iastate.edu
Order of Authors:	Tuyen Le
	H. David Jeong, Ph.D.
Suggested Reviewers:	
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Authors are required to attain permission to re-use content, figures, tables, charts, maps, and photographs for which the authors do not hold copyright. Figures created by the authors but previously published under copyright elsewhere may require permission. For more information see http://ascelibrary.org/doi/abs/10.1061/978	No

0784479018.ch03. All permissions must be uploaded as a permission file in PDF format. Are there any required permissions that have not yet been secured? If yes, please explain in the comment box.	
ASCE does not review manuscripts that are being considered elsewhere to include other ASCE Journals and all conference proceedings. Is the article or parts of it being considered for any other publication? If your answer is yes, please explain in the comments box below.	No
Is this article or parts of it already published in print or online in any language? ASCE does not review content already published (see next questions for conference papers and posted theses/dissertations). If your answer is yes, please explain in the comments box below.	No
Has this paper or parts of it been published as a conference proceeding? A conference proceeding may be reviewed for publication only if it has been significantly revised and contains 50% new content. Any content overlap should be reworded and/or properly referenced. If your answer is yes, please explain in the comments box below and be prepared to provide the conference paper.	No
ASCE allows submissions of papers that are based on theses and dissertations so long as the paper has been modified to fit the journal page limits, format, and tailored for the audience. ASCE will consider such papers even if the thesis or dissertation has been posted online provided that the degree-granting institution requires that the thesis or dissertation be posted. Is this paper a derivative of a thesis or dissertation posted or about to be posted on the Internet? If yes, please provide the URL or DOI permalink in the comment box below.	No
Each submission to ASCE must stand on its own and represent significant new information, which may include disproving the work of others. While it is acceptable to build upon one's own work or replicate other's work, it is not appropriate to fragment the research to maximize the number of manuscripts or to submit	No

papers that represent very small incremental changes. ASCE may use tools such as CrossCheck, Duplicate Submission Checks, and Google Scholar to verify that submissions are novel. Does the manuscript constitute incremental work (i.e. restating raw data, models, or conclusions from a previously published study)?	
Authors are expected to present their papers within the page limitations described in <u>Publishing in ASCE Journals: A Guide for Authors</u>. Technical papers and Case Studies must not exceed 30 double-spaced manuscript pages, including all figures and tables. Technical notes must not exceed 7 double-spaced manuscript pages. Papers that exceed the limits must be justified. Grossly over-length papers may be returned without review. Does this paper exceed the ASCE length limitations? If yes, please provide justification in the comments box below.	No
All authors listed on the manuscript must have contributed to the study and must approve the current version of the manuscript. Are there any authors on the paper that do not meet these criteria? If the answer is yes, please explain in the comments.	No
Was this paper previously declined or withdrawn from this or another ASCE journal? If so, please provide the previous manuscript number and explain what you have changed in this current version in the comments box below. You may upload a separate response to reviewers if your comments are extensive.	No
Companion manuscripts are discouraged as all papers published must be able to stand on their own. Justification must be provided to the editor if an author feels as though the work must be presented in two parts and published simultaneously. There is no guarantee that companions will be reviewed by the same reviewers, which complicates the review process, increases the risk for rejection and potentially lengthens the review time. If this is a companion paper, please indicate the part number and provide the title, authors and manuscript number (if available) for the companion papers along with your detailed justification for the editor in the comments box below. If there is no justification provided, or if there is	

insufficient justification, the papers will be returned without review.	
If this manuscript is intended as part of a Special Issue or Collection, please provide the Special Collection title and name of the guest editor in the comments box below.	
Recognizing that science and engineering are best served when data aremade available during the review and discussion of manuscripts andjournal articles, and to allow others to replicate and build on workpublished in ASCE journals, all reasonable requests by reviewers formaterials, data, and associated protocols must be fulfilled. If you are restricted from sharing your data and materials, please explain below.	
Papers published in ASCE Journals must make a contribution to the core body of knowledge and to the advancement of the field. Authors must consider how their new knowledge and/or innovations add value to the state of the art and/or state of the practice. Please outline the specific contributions of this research in the comments box.	The data inconsistency among project phases, stakeholders, or geographic regions (counties, states, etc.) is a major barrier to digital exchange of life cycle project data in the civil infrastructure industry. However, research to address the issue of terminology inconsistency in the construction industry has been very limited. This research has attempted fill that gap by proposing a novel methodology for 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 in which semantic relations (synonyms, hyponyms, hypernyms, attributes, etc.) among terms are explicitly represented. The InfraLex dictionary enables computer systems to understand the meaning of a piece of data; hence, data mismatch when integrating isolated datasets will be eliminated. The study is, therefore, intended to make contributions to facilitating seamless data exchange of digital data across the life-cycle of an infrastructure project. With explicit explanations of technical terms, InfraLex is also expected to become a fundamental resource for natural language processing studies in the civil infrastructure domain.
The flat fee for including color figures in print is $800, regardless of the number of color figures. There is no fee for online only color figures. If you decide to not print figures in color, please ensure that the color figures will also make sense when printed in black-and-white, and remove any reference to color in the text. Only one file is accepted for each figure. Do you intend to pay to include color figures in print? If yes, please indicate which figures in the comments box.	No
If there is anything else you wish to communicate to the editor of the journal, please do so in this box.	

NLP-based approach to classifying heterogeneous terms for unambiguous exchange of roadway data

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

ABSTRACT

The inconsistency of data terminology due to the fragmented nature of the highway industry has imposed big challenges on integrating digital data from distinct sources. The issue of semantic heterogeneity may lead to the lack of common understanding of the same data between the sender and receiver. Explicit semantic relations among terms in digital dictionaries, such as ontologies can enable the meaning of a roadway concept name to be transparent and unambiguously understood by computer systems. However, due to the lack of an effective automated method, current practices of identifying these relations hardly rely on a manual process of knowledge acquisition from domain experts or text documents which is laborious and time-consuming. This paper presents a novel methodology that leverages recent advances in Natural Language Processing (NLP) techniques to extract English-American roadway terms used in different government agencies and their semantic relations from roadway design manuals and specifications. The proposed method includes the following three stages: (1) implementing NLP techniques to detect commonly used technical terms from the highway corpus; (2) utilizing machine learning to learn the semantic similarity among roadway terms using their context data in the corpus; and (3) developing a classification algorithm to identify semantic relation types among technical terms. The key merit in this technique is the automated identification of semantic relations among heterogeneous roadway terms from design guidebooks without reliance on other existing hand-coded

[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.

semantic resources. The proposed methodology was evaluated by conducting an experiment comparing the automatically-identified synonyms by the proposed system with a human-constructed golden standard dataset obtained from Wikipedia. The result shows that the proposed model achieves a precision of over 80 percent.

Keywords: Roadway Data, Data Sharing, Semantic Interoperability, Semantic Relation, Natural Language Processing, Vector Space Model

INTRODUCTION

The implementation of advanced, and computerized technologies such as 3D modeling and Geographic Information System (GIS) throughout the life cycle of a highway project has allowed a large portion of project data to be available in a 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. The interoperability issue has been widely recognized as a key obstacle blocking the flow of digital data throughout the entire project life cycle. The inadequate interoperability cost is estimated to be over $15.8 billion per year in the U.S. capital facilities industry as reported by the National Institute of Standard and Technology (NIST); and the largest cost item is the laborious work for finding, verifying, and transferring facility and project information into a useful format during the operation and maintenance stage (Gallaher et al. 2004). 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 major domains in the construction industry, has not yet successfully facilitated a high degree of interoperability (Lefler 2014); huge cost savings would be achieved if roadway data is seamlessly shared across project phases and among state and local agencies.

Semantic interoperability, which relates to the issue whereby two computer systems may not share a common understanding of a specific 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 differs between phases, stakeholders, or geographic regions (counties, states, etc.). Retrieving 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, *roadway type* can either mean the classification of roadways by material or function. Walton et al. (2015) suggests the following three reasons for semantic heterogeneity among transportation databases: (1) isolation in definitions among separate sources, (2) temporary of definitions and (3) variety of data collection methods. Synonymy, in contrast, is associated with a set of different terms used to present the same concept. For instance, 'profile', 'crest', 'grade-line' and 'vertical alignment' are equivalent terms of the *longitudinal centerline* of a roadway. Under these situations, simply mapping of data names will likely lead to a failure of data extraction, or use of wrong data. Thus, addressing the semantic inconsistency issue becomes crucial to ensure a common understanding of the same dataset among software applications and guarantee a proper integration of data from multiple sources.

Terminology transparency through digital dictionaries like glossaries, taxonomies, ontologies and data dictionaries is identified as a driver of semantic interoperability (Ouksel and Sheth 1999). Although a plethora of semantic resources have been introduced to the highway sector; as shown in the literature review, their coverages of concepts are still inadequate and the inclusion of multiple names for the same concept is limited. This is because of the reliance on a tedious and time-consuming approach which requires developers to manually gather and translate knowledge from domain experts or text documents into a machine-readable format. Thus, there is a need for computer-aided methods to remove this knowledge acquisition bottleneck (Mounce et al. 2010), so that digital dictionaries can be quickly constructed to meet a specific need and to be able to keep up with the growth of terms due to rapid applications of new technologies and knowledge.

Recent achievements in accuracy and processing time of advanced Natural Language

Processing (NLP) techniques have driven text mining and cognitive recognition research to a new era. There is a rich set of NLP tools that can support various text processing tasks ranging from basic grammar analyses of individual words (Toutanova et al. 2003; Cunningham et al. 2002), and their dependencies (Chen and Manning 2014), to deep learning of meanings (Mikolov et al. 2013; Pennington et al. 2014). These NLP advances offer numerous potentials for the construction industry where most of the domain knowledge resources are in text documents (e.g., design guidelines, specifications). The implementation of NLP will allow for a fast translation of the domain knowledge into a computer-readable format which is required for machine-to-machine based data exchange.

This paper presents an NLP-based automated approach to gather commonly used American-English roadway terms in different highway agencies and classify the semantic relations among these heterogeneous terms. This study leverages NLP and machine learning to extract meanings of terms by analyzing their statistical data of context words in various state design manuals. The semantics of roadway terms are represented as vectors in a high dimensional coordinate system in which the semantic similarity among terms is quantifiable. The proposed methodology also includes a new classification algorithm that utilizes syntactic rules and cluster analysis to categorize related terms into three different groups that are synonyms, hyponyms, and attributes. A Java package built upon the proposed method and several datasets resulting from the study can be found at https://github.com/tuyenbk/mvdgenerator.

BACKGROUND

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. Some of the major applications of NLP include language translation, information extraction, opinion mining (Cambria and White 2014). These applications are embodied by a rich set of NLP techniques ranging from grammar processing such as Tokenization (breaking a sentence into individual tokens)

(Webster and Kit 1992; Zhao and Kit 2011), Part-of-Speech (POS) tagging (assigning tags, adjective, noun, verb, etc. to each token of a sentence) (Toutanova et al. 2003; Cunningham et al. 2002), and Dependency parser (identifying relationships between linguistic units) (Chen and Manning 2014), to the semantic level, for instance word sense disambiguation (Lesk 1986; Yarowsky 1995; Navigli 2009). NLP methods can be classified into two main groups: (1) rule-based and (2) machine-learning (ML) based methods. Rule-based systems, which rely solely on hand-coded syntax rules, are not able to fully cover all human rules (Marcus 1995); and their performances, therefore, are 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 quickly 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 semantic similarity, which is one of the important NLP-related research topics, aims to determine how much two linguistic units (e.g., words, phrases, sentences, concepts) are semantically alike. For example, a *railway* might be more similar to a *roadway* than to *train*. The state-of-the-art methodology for this task can be divided into two categories that are (1) thesaurus-based methods and (2) vector space models (VSM) (Harispe et al. 2013). The former approach relies on a hand-coded digital dictionary (e.g., WordNet) that formally structures terms through a network of semantic relations. Computational platforms (e.g., information retrieval) built upon such dictionaries measure the semantic similarity between a given pair of words by computing the length of their connecting path in the hierarchy. 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, assesses the meanings of words or phrases by analyzing their occurrence frequencies 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 can be represented by the Euclidean distance between the corresponding points (Erk 2012). The conventional method to construct a VSM is to use the 'word-context' matrix which shows how frequent a word is the context of one another in a given text corpus. These raw data of frequencies are used to estimate the co-occurrence probabilities. This statistical process results in a new matrix in which each row is a vector representation. Pointwise Mutual Information (PMI) (Church and Hanks 1990) or it's variant, Positive PMI (PPMI) is a popular method to calculate the co-occurrence probabilities. A more advanced approach uses machine learning to train the representation vectors of terms. One example of this line of methodology is the Skip-gram neural network model (Mikolov et al. 2013) which aims to predict the context words of a given input word. The training objective is to minimize the overall error between the predicted and the actual context vectors. Glove (Pennington et al. 2014), an alternative machine learning model for building VSM, trains on the global 'word-context' matrix with the objective that the probability of co-occurrence between two words equals the dot product of their vector representations. The major difference between these two models is that Skip-Gram model trains the local context data within a context window, Glove trains on the global co-occurrence statistics. There are contradictive recommendations on the wining model in the literature. The authors of Glove suggested that their model out-performs Skip-Gram and others in the state of the art. However, a number of independent benchmarking experiments have consistently indicated the outperformance of the Skip-gram model to its alternatives.

For example, a comparative study conducted by Levy et al. (2015) on the accuracy in various tasks and golden standards reveals that Skip-gram outperforms Glove in every experiment and is the winner in most of the tasks, especially on the WordSim Similarity dataset. Among these tasks, the best precision of Skip-gram is .793, while PPMI and Glove achieve the highest score of .755 and .725 respectively. The out-performance of Mikolov's model on the similarity task is confirmed in another benchmarking study (Hill et al. 2015) where this model is also found as the winner in most of the tests.

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

Related studies

A popular solution to semantic interoperability is to develop taxonomies, ontologies or other forms of digital dictionaries that can provide machine-readable definitions of domain concepts. A plethora of such semantic resources have been developed for the highway industry. However, conventional development methods require significant human efforts on knowledge retrieval, and ontology construction and validation. The pioneer in this line of research is the e-Cognos ontology (Wetherill et al. 2002; Lima et al. 2005) which formulates the execution process of a construction project as an explicitly interactive network of the following principal concepts: Actor, Resources, Products, Processes and Technical Topics. The ontology developers of this project reviewed existing taxonomies (BS61000, UniClass, IFC) and construction specific documents, and interacted with the end users to identify

relevant concepts and their semantic relations. Industry experts were invited to validate e-Cognos through questionnaires on concept names and relations. Since the introduction of the high-level ontology of e-Cognos, a plenty of ontologies have been built for various aspects of a highway project, for instance, highway construction taxonomy (El-Diraby and Kashif 2005; El-Diraby et al. 2005), freight ontology (Seedah et al. 2015a), and the ontology of urban infrastructure products (Osman and Ei-Diraby 2006). Like the e-Cognos project, these studies also relied on domain experts for constructing their semantic products. The limitation regarding time and labor costs of the ad-hoc traditional methodology has created a bottleneck to the progress in enabling semantic interoperability. In addition, the existing ontologies primarily focus on the description of concepts, the heterogeneity of concept names is usually neglected. Therefore, research is needed not only to automate the process of formulating domain concepts but also to incorporate term heterogeneity into ontologies.

Another strategy for semantic interoperability targets at the heterogeneity of concept names rather the concept description as in an ontology model. A few frameworks to assist practitioners in precisely mapping data labels from heterogeneous sources have been introduced for various construction sectors. In the building sector, buildingSMART proposed a novel framework, namely IFD (International Framework for Dictionaries) (ISO 12006-3) for developing a multilingual data schema in which each concept can have multiple names in different languages. With IFD, the identity of a concept is defined by a Global Unique ID (GUID) rather than its name; hence an IFD-based data exchange mechanism is able to eliminate the semantic mismatches due to the name inconsistency (IFD Library Group 2008; Hezik 2008). The buildingSMART data dictionary (bSDD) (buildingSMART 2016) is the first digital library of building concepts that is crafted in the IFD structure. Each concept in bSDD consists a set of synonymy names not only in English but also in computer-coded languages (e.g., IFC) and in other human languages (e.g., French, Norwegian). Therefore, a complete bSDD would enable digital data in regardless of languages to be sharable and unambiguously reusable. Yet, its size remains limited as the identification of these sets

of synonyms is labor and time extensive. In the transportation sector, there has been a shortage of research efforts targeting the heterogeneity of data names at the database level until recently. Seedah et al. (2015b) proposed a role-based classification schema (RBCS) to classify data in freight databases. RBCS defines nine distinct groups of roles that are time (year, month), place (city name, population), commodity (liquid, value), link (roadway name, width), mode (truck, rail), industry (company name, sales), event (accident, number of fatalities), and human (officer, driver age). The authors argue that once the data elements across separate databases are categorized using this standard system, it becomes easier for practitioners to identify the semantic relatedness in their definitions. However, even if RBCS is successfully applied to all freight databases, identifying the exact type of relation (synonym, functional relation) between two data elements in the same category is still a challenging task.

In attempts to reduce laborious work on defining concepts, a few researchers have sought to propose semi-automated and automated methods for identifying semantic relations among technical terms. Abuzir and Abuzir (2002) developed the ThesWB system which utilizes hand-coded syntax patterns to detect lexical relations between civil engineering terms from HTML web pages. The performance of ThesWB was not reported, but it is not likely to be high since rule-based approaches are repeatedly criticized for not being able to capture all the variant ways to present relations among terms in natural language (Marcus 1995; Navigli and Velardi 2010). Rezgui (2007) suggested a more sophisticated approach that is based the statistics of word occurrence rather than predefined rules to extract potential pairs of related terms from domain text documents. This method implements TF-IDF to evaluate the importance degree of a keyword to the examined domain; and analyzes the co-occurrence frequencies using Metric Clusters to assess the potentiality that exists a semantic relation within a given pair of important keywords. These potential relationships are then validated and categorized by domain experts. Since the method detects relations between words occurring in the same sentence, equivalent terms which are used interchangeably

could not be captured. In another study to identify semantic relations, Zhang and El-Gohary (2016) proposed a fully automated methodology for both tasks of retrieving related candidate and classifying the relations. This algorithm was reported to achieve an average precision of nearly 90 percent in the relation classification task. However, the algorithm identifies potentially related concepts based on the pre-defined lexical relations provided in WordNet, a generic lexicon that lacks concepts in many construction sectors including the civil infrastructure, it would not be scalable well on matching terms in these domains.

As shown in the literature review, there are numerous research efforts in developing ontologies for the highway sector. However, the existing ontologies are mainly hand-coded through the manual processes of knowledge acquisition and translation into a digital format. This ad-hoc approach has created a bottleneck in facilitating the semantic interoperability for the whole industry and as a result, semantic resources for many aspects of a project are still not available. A few efforts have been made to automate the process of constructing or extending existing semantic resources. The most rigorous methodology in the state-of-the-art is the one developed by Zhang and El-Gohary (2016) that is fully-automated with high accuracy. One limitation of this algorithm is the reliance on an existing semantic resource; it, therefore, would not be applicable to such a domain like the infrastructure that is out of the vocabulary scope. Thus, there is a need for an automated approach that can not only allow for a fast development of highway lexicons but also remove the dependence on other existing semantic models.

PROPOSED METHODOLOGY TO AUTOMATED CLASSIFICATION OF ROADWAY TERMS

The goal of this research is to propose an NLP-based methodology that can automate the process of extracting roadway technical terms and their semantic relations from American-English roadway documents. As shown in Figure 1, the proposed methodology consists of three major modules that are to: (1) utilize NLP techniques to extract multi-word roadway technical terms from a collected text corpus, (2) train the data obtained form the text corpus

using the Skip-gram neural network model (Mikolov et al. 2013) to develop a Roadway Vector Space Model (Rd-VSM) that presents the semantics of roadway terms, and (3) develop an algorithm integrating Rd-VSM and various linguistic patterns to classify relations among technical terms (synonyms, hyponyms and attributes). The below sections discuss these steps in detail.

Text corpus collection

In order to capture the heterogeneity of roadway terms, the authors collected a plethora of highway engineering manuals and guidelines from 30 State Departments of Transportation. The content of a written guidance document in the engineering field is commonly presented in various formats such as plain text, tables, and equations. Since the structures of 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 may slightly reduce the corpus size, and accordingly affects the training dataset; however, it is necessary since words in tables and equations are not organized in the formal structure of a sentence and therefore the NLP algorithm may extract unreal noun phrases. The final outcome of this phase is a plain text corpus consisting of nearly 16 million words. This dataset is utilized to extract multiple-word technical terms which are then trained and transformed into representation vectors.

Multi-word terms extraction

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). The meaning of a multi-word term may not be directly interpreted from the meanings of its constituents; therefore, it must be treated as an individual word. To meet that requirement, multi-word terms in the corpus need to be detected and replaced with connected blocks of their members. As mentioned, a multiple-word term must be an NP; thus, NPs will be good multi-word term candidates. To detect this type of term, the corpus is first scanned to search for NPs, of which the importance is then evaluated based on their

statistics of occurrence. The process of extracting multi-word terms is discussed in detail as follows.

Noun phrase extraction

This research implements the Apache OpenNLP package to find sequences of words that match pre-defined noun phrase patterns. Figure 2 illustrates how noun phrases are extracted from the corpus of highway technical documents. This 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., NN-noun, JJ-adjective, VB-verb, etc.) for each unit of the tokenized corpus obtained from the previous step. A full set of POS tags can be found in the Penn Treebank (Marcus et al. 1993).

iii Noun phrase detection: 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 to collect sequences matching those patterns. In addition, in order to reduce the discrimination between the syntactic variants of the same term, the collected NPs need to be normalized. The following discuss two types of syntactic variants considered and the proposed normalization methods.

- Type 1 - Plural forms, for example 'roadways' and 'roadway'. Stemming is a popular process to reduce words to their stems. Despite the fact that, none of the existing algorithms can completely eliminate the errors of over and under stemming, they are good enough to not degrade the overall performance of NLP applications (Jivani et al. 2011). This study implements the Pling stemmer (Suchanek et al. 2006), which stems an English noun to its singular form, to normalize plural nouns in the corpus. One advantage of this algorithm is the

utilization of both syntactic rules and the vocabulary in a dictionary; hence the miss- or over-stemming errors that take off a true suffix can be reduced.

- Type 2 - Preposition noun phrases, for example 'type of roadway' and 'roadway type'. 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'.

The first column in Table 2 represents several examples of the NP bag retrieved from this phase. Since an NP is not certainly a technical term, those that are clearly unlikely to be a term should be excluded from the candidate list. Occurrence frequency is a key indicator for the importance of a candidate as a technical term tends to repeatedly occur in domain text documents. To eliminate 'bad' candidates, a threshold of frequency can be applied. If users choose a high threshold, rare terms would not be captured. This issue can be addressed when the corpus size is extended. In our experiment, with a frequency threshold of 2, the final list of NPs consists of 112,024 items; and it drops to 8,922 when a threshold of 50 is used. Since this research aims at common technical terms, the authors used a threshold of 50 to remove possibly meaningless term candidates.

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 an 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 outperforms other methods on the Wikipedia corpus, and C-Value is 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 an 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.

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

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 automatically remove candidates that are unlikely to be real terms, a threshold C-value can be used. However, doing this may eliminate the real terms that appear in the bottom due to their low frequencies. Manual evaluation of the entire candidate list would avoid the removal of real terms with low C-values. To minimize both laborious work and

the number of true terms wrongly discarded, the authors suggest the following method to identify the threshold value. The ranked list of candidates is divided into groups of around 200 items. A graduate student with a civil engineering background was asked to utilize a bottom-up approach to evaluate group by group and stop at which the percentage of actual terms achieved 80 percent. Users can choose a higher percentage limit in cases where the accuracy is critical. This will increase manual evaluation effort. Table 2 illustrates the evaluation results for several excerpts of the extracted term candidates. The precision values, which represent the percentages of real terms in these groups, are presented in Figure 3. As shown in the figure, precision values are less than 80 percent for groups with c-values less than 50. This value is set as the threshold for the acceptance of term candidates. The final selected list is comprised of nearly 8,000 multi-word roadway technical terms.

Construction of term space model

This step aims at converting the vocabulary in the roadway corpus into a vector space model, namely Rd-VSM. Skip-gram (Mikolov et al. 2013), which is an un-supervised machine model, is 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 that appear around the input unit in the corpus. In order to collect this training dataset, the tokenized and stemmed 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 as single words. To fulfill that requirement, the white spaces within a multi-word term are replaced with minus (-) symbols to connect its individual words into a single unit. 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 window size of 5 will be the following word set {bike, lane, width, on, a, with, no, curb, and, gutter}. Any context word that is in the stop list (a list that contains frequent words in English such as 'a', 'an', and 'the' that have little meaning) will be neglected from the context set. In this example, the adjusted context set of 'roadway' is {bike, lane, width, curb, gutter}.

"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 illustrates 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 element at the index of the word in the vocabulary is set one, and all other items are zero. For example, the one-hot vector of k^{th} word in the vocabulary with the size of V will be $\{x_1 = 0, x_2 = 0, ..., x_k = 1, ...x_V = 0\}$. The outcome of this machine learning process is a set of word representation vectors in an N-dimension coordinate system. The similarity among these vectors represents the similarity in context between the corresponding words. The bullets below explain how the predicted context vector of k^{th} word is computed using the parameter matrices resulted from the learning process. As we can see, the similarity between two predicted context vectors depends only on the similarity between their corresponding input representation vectors; thus, these vectors are used to represent the semantics of words.

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

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

The learning 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, Word2Vsec 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 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. A reasonable figuration for this parameter is from tens to hundreds (Radim 2014). The final major parameter, *context window size*, decides how many context words to be considered. Google recommends a 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 Rd-VSM vector space model derived from the training process when the parameters, *frequency threshold*, *hidden layer size* and *window size* are set 50, 300 and 10 respectively. In this model, each word in the highway corpus is represented as a vector in a high dimensional space. Since the representation vectors are in a multi-dimensional space; to present the space in 2D graph, PCA (Principle Component Analysis) is used to reduce the dimension size to two.

The similarity between terms in the Rd-VSM model can be measured by the angle be-

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

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

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

Where: n is the hidden layer size.

Semantic relation classification

The purpose of this module is to design an algorithm for automated classification of the semantic relations among the roadway technical terms. This study considers three core relation types of a semantic resource that are: synonym (meaning equivalence), hypernym-hyponym (also known as IS-A or parent-child relation), attribute (concept property) (Jiang and Conrath 1997; Lee et al. 2013). The following describes the fundamental logic behind the designed algorithm. Two terms that relate to each other through these semantic relations would have a high similarity score. Therefore, the top nearest terms resulted from Rd-VSM would be a great starting point for detecting relations between technical terms. For example, in the list of the nearest terms of 'roadway' (see Table 4), true synonyms are 'highway' (rank 1), 'traveled-way' (2) and 'road' (4); attributes include 'roadway-section' (3), 'roadway-shoulder' (12); and 'adjacent-roadway' (7) and 'undivided' (37) are hyponyms which show different types of roadway.

Algorithm 1 shows the designed 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. The algorithm first detects terms belonging to the first two categories using linguistic pat-

terns, and employs cluster analysis for the last group.

Attributes and hyponyms

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 its *Noun2* component.

Synonyms

After the words in the first two categorized are classified, the remained nearest words will fall into the third group. However, some of them may have far or even no relation with the target word. In order to address this issue, this framework employs the K-mean clustering algorithm (MacQueen 1967) to split the remained list into multiple layers based on the similarity score. Those terms in the last layers are unlikely to be synonyms; and thus, are removed from the classified list. Only the terms in the top cluster are kept and categorized as synonyms.

By the end of the synonym recognition phase, the algorithm will generate a list of classified nearest terms for a given input word. Table 6 shows one example of the output generated by the algorithm.

PERFORMANCE EVALUATION

This section presents a performance evaluation of the proposed system 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 from a Wikipedia transportation glossary (Wikipedia 2016). The developed algorithm was employed to find the synonym for a given input term. The automatically identified synonym is the nearest word in the synonym 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, or "false" if it does not. The answer will be "N/A" if the target

Algorithm 1 Semantic relation classification algorithm

Inputs: term t, list of nearest terms N, list of multi-word terms F
Output:: Classified list of terms C
procedure TERM CLASSIFICATION PROCEDURE
    $Att \leftarrow$ list of attributes
    $Hyp \leftarrow$ list of hyponyms
    $Syn \leftarrow$ list of synonyms
    $w \leftarrow null$
    for all $n \in N$ **do**
        if n contains t **then**
            $w \leftarrow n$
        else
            for all $f \in F$ **do**
                if f contains both n and t **then**
                $w \leftarrow f$
                Break for
        if w matches *Attribute pattern* **then**
            add w to Att
        else if w matches *Hyponym pattern* **then**
            add w to Hyp
        else
            add w to Syn
    Cluster Syn and discard low relevant terms

term is out of the model vocabulary. 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.

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

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

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

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 the configuration of these parameters one by one to evaluate their effects to the model performance. While changing a certain parameter, other parameters are kept unchanged compared to their values in the base model. As presented in the table, the model performance is not significantly sensitive to the changes of training parameters. 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 changes of other parameters did not improve the performance. Especially, the increase of frequency threshold value from 50 to 75 has negative impact to all measures. This result confirms the reasonable selection of the frequency threshold to eliminate unlikely term candidates in the NP extraction phase.

The proposed model was also compared with the generic WordNet database. Table 8 presents the comparison of performance between the proposed framework (with the 50-100-10 setting) and WordNet. As shown, the present system 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 roadway vocabulary.

DISCUSSIONS

This paper proposes an NLP based methodology to assist professionals in extracting roadway terms and their semantic relations from text documents. A key contribution to the body of knowledge is the novel framework with a new algorithm that allows for automated detection of technical terms and their relations without reliance on existing hand-coded dictionaries as used by previous researchers such as Zhang and El-Gohary (2016). The present framework is not to completely eliminate human involvement, but is expected to significantly reduce manual efforts and become an enabling tool that can help researchers in the highway domain quickly develop supporting ontologies and other forms of semantic resources for their specific use cases. With respect to the facilitating semantic interoperability for the infrastructure sector, the findings of this study would accelerate the process of removing the current bottleneck in extensive machine readable dictionaries which are required for an

unambiguous data sharing, integration or exchange.

The semantic similarity model and the relation classification algorithm developed in this study are also expected to become fundamental resources for a variety of NLP related studies in the highway domain. NLP based platforms can utilize these resources 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. Information retrieval systems can use the semantic relations provided by the algorithm to classify project documents by relevant topics by analyzing the relatedness between the index keywords in those documents. In addition, questionnaire designers can utilize the system 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 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 synonym detection function 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. Moreover, this study can make fundamentally transform to the way human interacts with a 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.

The current study has a number of limitations. 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 size by adding more documents from other state agencies and disciplines (e.g., survey, construction, operation and maintenance) would enable more technical terms

to be covered in the vector space model; and consequently, the recall would be improved. Secondly, the number of semantic relation categories is limited to only three types of relations that are attributes, hyponyms and synonyms. There are other important semantic relations that are not considered such as hypernyms, siblings, functional associations, etc. Including these relations would reduce incorrect synonym matching, which will enhance the precision value, for those cases that a word does not have any equivalent term. Third, this study only targets at the synonymy issue, the issue of polysemy is not yet addressed. Further research is needed to detect different senses of a roadway term. One potential solution is to apply cluster analysis on the instances of context to determine the possibility that a term would have multiple meanings.

CONCLUSIONS

Data manipulation from multiple sources is a challenging task in highway asset management due to the inconsistency of data format and terminology. The contribution of this study is an NLP-based approach to automated classification of semantic relations among roadway technical terms based on their word occurrences in domain text documents. 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 30 State DOTs across the U.S. Machine learning is used to train the semantic similarity between technical terms. An algorithm is designed to classify the nearest terms resulted from the semantic similarity model into distinct groups according to their lexical relationships.

The developed system 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 a significant improvement has been made in comparison with an existing thesaurus database, the overall performance is not relatively high. This might be due to the limited 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 proposed automated methodology for detecting semantic relations from manuals is expected to significantly reduce human efforts in developing a semantic resource for a specific use case within the highway domain and become an enabler for semantic interoperability in this domain. The research also opens a new gate for computational tools regarding natural language processing in the highway sector. The developed system would enable computer systems to understand terms and consequently transform the way human interacts with a 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>.

buildingSMART (2016). "buildingsmart data dictionary, <http://bsdd.buildingsmart.org/>. (Accessed: March 15, 2016).

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

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

Church, K. W. and Hanks, P. (1990). "Word association norms, mutual information, and lexicography." *Computational linguistics*, 16(1), 22–29.

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.

El-Diraby, T. and Kashif, K. (2005). "Distributed ontology architecture for knowledge management in highway construction." *Journal of Construction Engineering and Management*, 131(5), 591–603.

El-Diraby, T., Lima, C., and Feis, B. (2005). "Domain taxonomy for construction concepts: Toward a formal ontology for construction knowledge." *Journal of Computing in Civil 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.

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

IFD Library Group (2008). "Ifd library white paper. Accessed: 2015-07-06.

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

Jivani, A. G. et al. (2011). "A comparative study of stemming algorithms." *Int. J. Comp. Tech. Appl*, 2(6), 1930–1938.

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*.

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). "Nchrp synthesis 458: Roadway safety data interoperability between local and state agencies." *Report no.*, Transportation Research Board.

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.

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*.

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.

Navigli, R. and Velardi, P. (2010). "Learning word-class lattices for definition and hypernym extraction." *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 1318–1327.

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.

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

Osman, H. and Ei-Diraby, T. (2006). "Ontological modeling of infrastructure products and related concepts." *Transportation Research Record: Journal of the Transportation Research Board*, 1984(-1), 159–167.

Ouksel, A. M. and Sheth, A. (1999). "Semantic interoperability in global information systems." *ACM Sigmod Record*, 28(1), 5–12.

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

Radim, R. (2014). "Word2vec tutorial, <http://rare-technologies.com/word2vec-tutorial/>.

Rezgui, Y. (2007). "Text-based domain ontology building using tf-idf and metric clusters techniques." *The Knowledge Engineering Review*, 22(04), 379–403.

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

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

Seedah, D. P., Choubassi, C., and Leite, F. (2015a). "Ontology for querying heterogeneous data sources in freight transportation." *Journal of Computing in Civil Engineering*, 04015069.

Seedah, D. P., Sankaran, B., and O'Brien, W. J. (2015b). "Approach to classifying freight data elements across multiple data sources." *Transportation Research Record: Journal of the Transportation Research Board*, (2529), 56–65.

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

Suchanek, F. M., Ifrim, G., and Weikum, G. (2006). "Leila: Learning to extract information by linguistic analysis." *Proceedings of the 2nd Workshop on Ontology Learning and*

Population: Bridging the Gap between Text and Knowledge, 18–25.

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." *Journal of Information Technology in Construction (ITCon)*, 7, 183–196.

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. (2016). "Extending building information models semiautomatically using semantic natural language processing techniques." *Journal of Computing in Civil Engineering*, C4016004.

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.

List of Figures

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'.	

TABLE 2: Excerpts of the extracted candidate terms

Term candidate	**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

TABLE 3: Skip-gram model parameters

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

TABLE 4: Examples of top nearest words

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

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 classified list of 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

TABLE 7: Performance of the synonym matching task with various training settings

Parameter changed	**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 proposed system

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

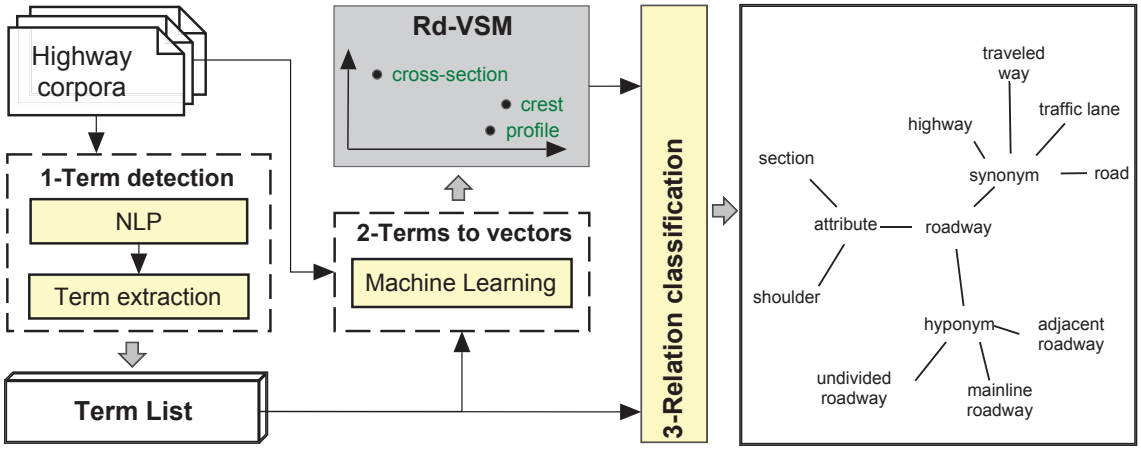

FIG. 1: Overview of the proposed methodology

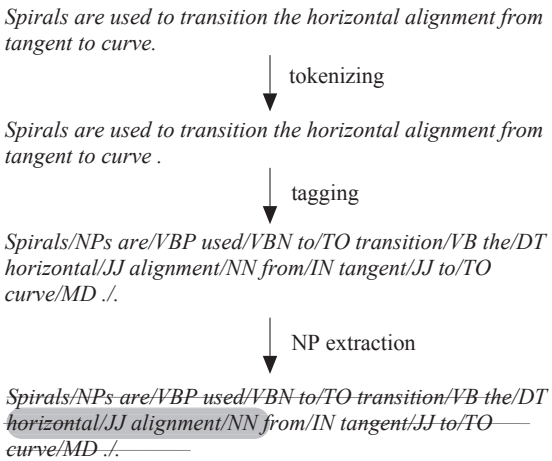

FIG. 2: Linguistic processing procedure to detect NPs

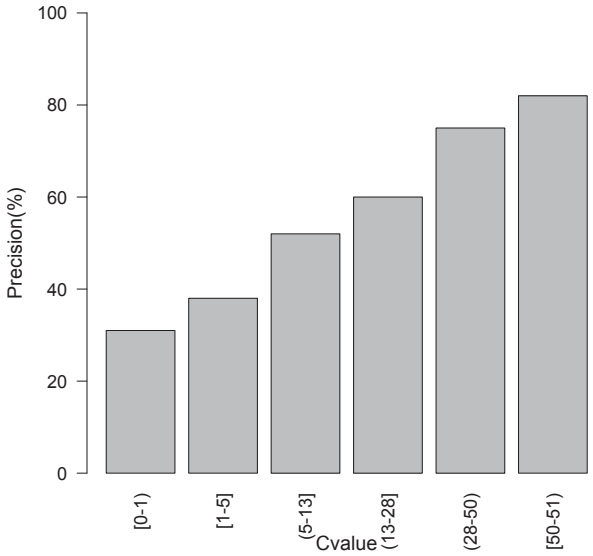

FIG. 3: Multi-word term extraction evaluation

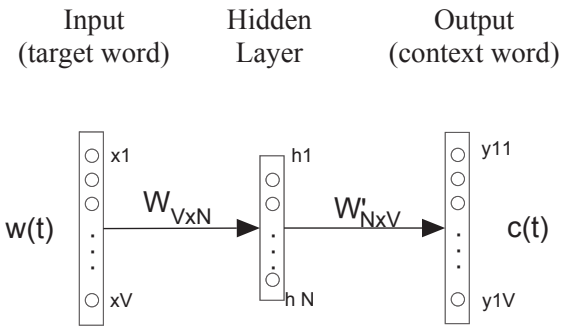

FIG. 4: Skip-gram model

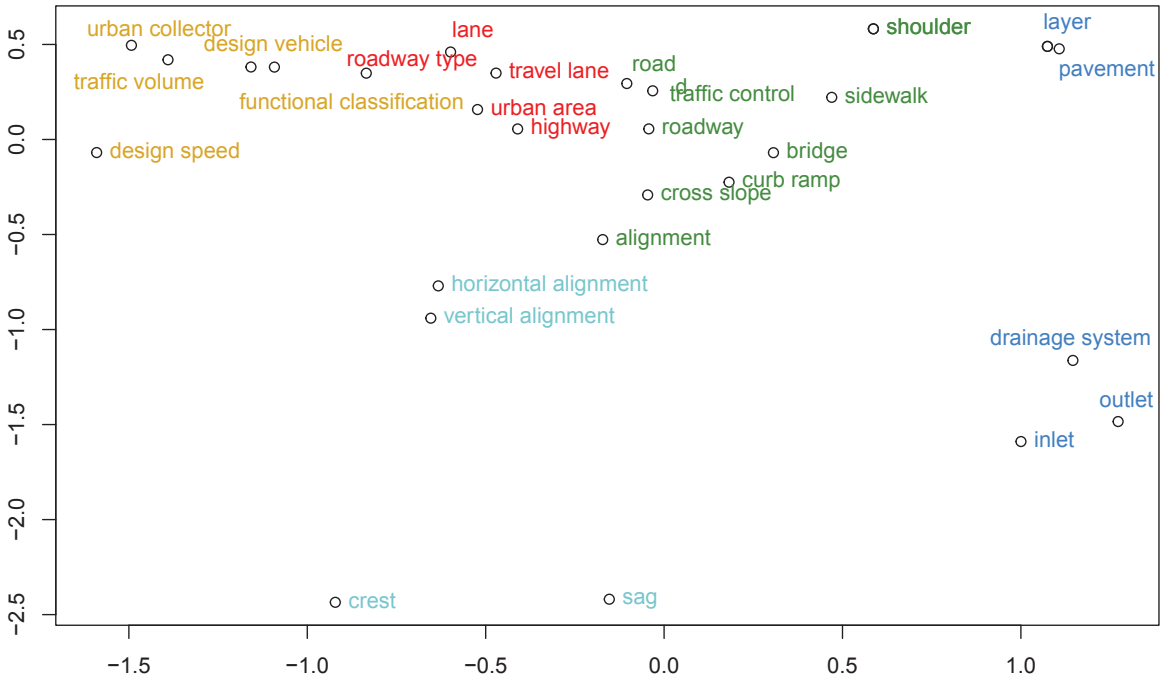

FIG. 5: Roadway term space model (Rd-VSM)

ASCE Journals Sizing Worksheet

Please complete this form for all new manuscripts

September 12, 2016

This worksheet will automatically calculate the total number of printed pages your article will occupy in the journal.

Please fill in all fields in green below. If you do not know your Manuscript Number, you may leave that field blank.

Length Limits:

Technical Paper/Case Study = 8 pgs.	**Forum = 4 pgs.**
Technical Note = 3 pgs.	**Discussion/Closure = 2 pgs.**

Manuscript number:	CPENG-1952
Journal name:	Journal of Computing in Civil Engineering
Corresponding author name:	H. David Jeong
Email address:	djeong@iastate.edu

Information on the maximum allowed length for each article type can be found online at: http://www.asce.org/Content.aspx?id=29559

Number of pages in your manuscript:	29
Number of figure pages:	5
Number of table pages:	8
Estimated article pages:	**10**

- Please include figure captions when indicating the size of your manuscript.

- Manuscripts should use 12 pt. font, double spaced, with 1 inch margins.

Note: The total displayed above is only an estimate. Final page count will depend on a number of factors, including the size of your figures and tables, and the number of display equations in your manuscript.

Additional author resources can be found online using the ASCE Author Guide located at: http://www.asce.org/Content.aspx?id=18107

ASCE Authorship, Originality, and Copyright Transfer Agreement

Publication Title: Journal of computing in civil engineering

Manuscript Title: NLP-based approach to classifying heterogeneous terms for unambiguous exchange of roadway data

Author(s) – **Names, postal addresses, and e-mail addresses of all authors**

Tuyen Le, Ph.D. Candidate, Department of Civil, Construction and Environmental Engineering, Iowa State Uni- versity. Ames, IA 50011, United States. E-mail: ttle@iastate.edu.

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

I. Authorship Responsibility

To protect the integrity of authorship, only people who have significantly contributed to the research or project and manuscript preparation shall be listed as coauthors. The corresponding author attests to the fact that anyone named as a coauthor has seen the final version of the manuscript and has agreed to its submission for publication. Deceased persons who meet the criteria for coauthorship shall be included, with a footnote reporting date of death. No fictitious name shall be given as an author or coauthor. An author who submits a manuscript for publication accepts responsibility for having properly included all, and only, qualified coauthors.

I, the corresponding author, confirm that the authors listed on the manuscript are aware of their authorship status and qualify to be authors on the manuscript according to the guidelines above.

Hyungseok David Jeong — Hyungseok Jeong — 08/10/16

Print Name — Signature — Date

II. Originality of Content

ASCE respects the copyright ownership of other publishers. ASCE requires authors to obtain permission from the copyright holder to reproduce any material that (1) they did not create themselves and/or (2) has been previously published, to include the authors' own work for which copyright was transferred to an entity other than ASCE. Each author has a responsibility to identify materials that require permission by including a citation in the figure or table caption or in extracted text. Materials re-used from an open access repository or in the public domain must still include a citation and URL, if applicable. At the time of submission, authors must provide verification that the copyright owner will permit re-use by a commercial publisher in print and electronic forms with worldwide distribution. For Conference Proceeding manuscripts submitted through the ASCE online submission system, authors are asked to verify that they have permission to re-use content where applicable. Written permissions are not required at submission but must be provided to ASCE if requested. Regardless of acceptance, no manuscript or part of a manuscript will be published by ASCE without proper verification of all necessary permissions to re-use. ASCE accepts no responsibility for verifying permissions provided by the author. Any breach of copyright will result in retraction of the published manuscript.

I, the corresponding author, confirm that all of the content, figures (drawings, charts, photographs, etc.), and tables in the submitted work are either original work created by the authors listed on the manuscript or work for which permission to re-use has been obtained from the creator. For any figures, tables, or text blocks exceeding 100 words from a journal article or 500 words from a book, written permission from the copyright holder has been obtained and supplied with the submission.

Hyungseok David Jeong — Hyungseok Jeong — 08/10/16

Print name — Signature — Date

III. Copyright Transfer

ASCE requires that authors or their agents assign copyright to ASCE for all original content published by ASCE. The author(s) warrant(s) that the above-cited manuscript is the original work of the author(s) and has never been published in its present form.

The undersigned, with the consent of all authors, hereby transfers, to the extent that there is copyright to be transferred, the exclusive copyright interest in the above-cited manuscript (subsequently called the "work") in this and all subsequent editions of the work (to include closures and errata), and in derivatives, translations, or ancillaries, in English and in foreign translations, in all formats and media of expression now known or later developed, including electronic, to the American Society of Civil Engineers subject to the following:

- The undersigned author and all coauthors retain the right to revise, adapt, prepare derivative works, present orally, or distribute the work, provided that all such use is for the personal noncommercial benefit of the author(s) and is consistent with any prior contractual agreement between the undersigned and/or coauthors and their employer(s).
- No proprietary right other than copyright is claimed by ASCE.
- If the manuscript is not accepted for publication by ASCE or is withdrawn by the author prior to publication (online or in print), or if the author opts for open-access publishing during production (journals only), this transfer will be null and void.
- Authors may post a PDF of the ASCE-published version of their work on their employers' ***Intranet*** with password protection. The following statement must appear with the work: "This material may be downloaded for personal use only. Any other use requires prior permission of the American Society of Civil Engineers."
- Authors may post the ***final draft*** of their work on open, unrestricted Internet sites or deposit it in an institutional repository when the draft contains a link to the published version at www.ascelibrary.org. "Final draft" means the version submitted to ASCE after peer review and prior to copyediting or other ASCE production activities; it does not include the copyedited version, the page proof, a PDF, or full-text HTML of the published version.

Exceptions to the Copyright Transfer policy exist in the following circumstances. Check the appropriate box below to indicate whether you are claiming an exception:

☐ **U.S. GOVERNMENT EMPLOYEES:** Work prepared by U.S. Government employees in their official capacities is not subject to copyright in the United States. Such authors must place their work in the public domain, meaning that it can be freely copied, republished, or redistributed. In order for the work to be placed in the public domain, ALL AUTHORS must be official U.S. Government employees. If at least one author is not a U.S. Government employee, copyright must be transferred to ASCE by that author.

☐ **CROWN GOVERNMENT COPYRIGHT:** Whereby a work is prepared by officers of the Crown Government in their official capacities, the Crown Government reserves its own copyright under national law. If ALL AUTHORS on the manuscript are Crown Government employees, copyright cannot be transferred to ASCE; however, ASCE is given the following nonexclusive rights: (1) to use, print, and/or publish in any language and any format, print and electronic, the above-mentioned work or any part thereof, provided that the name of the author and the Crown Government affiliation is clearly indicated; (2) to grant the same rights to others to print or publish the work; and (3) to collect royalty fees. ALL AUTHORS must be official Crown Government employees in order to claim this exemption in its entirety. If at least one author is not a Crown Government employee, copyright must be transferred to ASCE by that author.

☐ **WORK-FOR-HIRE:** Privately employed authors who have prepared works in their official capacity as employees must also transfer copyright to ASCE; however, their employer retains the rights to revise, adapt, prepare derivative works, publish, reprint, reproduce, and distribute the work provided that such use is for the promotion of its business enterprise and does not imply the endorsement of ASCE. In this instance, an authorized agent from the authors' employer must sign the form below.

☐ **U.S. GOVERNMENT CONTRACTORS:** Work prepared by authors under a contract for the U.S. Government (e.g., U.S. Government labs) may or may not be subject to copyright transfer. Authors must refer to their contractor agreement. For works that qualify as U.S. Government works by a contractor, ASCE acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce this work for U.S. Government purposes only. This policy DOES NOT apply to work created with U.S. Government grants.

I, the corresponding author, acting with consent of all authors listed on the manuscript, hereby transfer copyright or claim exemption to transfer copyright of the work as indicated above to the American Society of Civil Engineers.

Hyungseok David Jeong

Print Name of Author or Agent

Hyung Jeong 09/10/16

Signature of Author of Agent Date

More information regarding the policies of ASCE can be found at http://www.asce.org/authorsandeditors

Response to Reviewers' Comments

Manuscript title: NLP-based approach to classifying heterogeneous terms for unambiguous exchange of roadway data

Manuscript #: CPENG-1952

The authors would like to thank the editor and reviewers for their time and effort in reviewing our manuscript. We hope that the revised manuscript is suitable for publication and we look forward to your response.

Editor's comments

1. "The manuscript has received two reviews. Both reviewers see the potential for publication, but have significant concerns (especially about the novelty and contribution). The authors are, thus, requested to revise the manuscript based on the reviews and resubmit it for another round of review. Please respond point-by-point to each review comment. In responding to a comment, please note any changes made in the manuscript and refer to the respective line numbers. When revising the manuscript, please better highlight your contribution, better justify the advancement in the body of knowledge (especially in comparison to recently published papers), and better describe the implications of the work and how successful the project is."

> *Response: The authors have revised the manuscript in accordance with the recommendations from the reviewers. Please see the responses below to each reviewer's comments. With respects to highlighting the research contribution, we have revised the manuscript with the aim of providing readers with a clearer discussions and justification for the research value. The following are examples of revised paragraphs discussing the contribution of the research throughout the revised manuscript.*
>
> ***Introduction section, page 3, line 65-75,*** *"Terminology transparency through digital dictionaries like glossaries, taxonomies, ontologies and data dictionaries is identified as a driver of semantic interoperability (Ouksel and Sheth 1999). Although a plethora of semantic resources have been introduced to the highway sector, their coverages of terms are still far inadequate and the inclusion of multiple names for the same concept is limited. This is because of the reliance on a tedious and time-consuming approach which requires developers to manually gather and translate knowledge from domain experts or text documents into a machine-readable format. Thus, there is a need for computer-aided methods to remove this knowledge acquisition bottleneck (Mounce et al. 2010), so that digital dictionaries can be quickly constructed to meet a specific need and to keep up with the growth of new terms due to rapid applications of new technologies and knowledge."*
>
> ***Related studies section, page 10, line 246-258,*** *"As shown in the literature review, there are numerous research efforts in developing ontologies for the highway sector. However, the existing ontologies are mainly hand-coded through manual processes of knowledge*

acquisition and formally describing them in a digital format. This ad-hoc approach has created a bottleneck in facilitating the semantic interoperability for the whole industry and as a result, semantic resources for many aspects of a project are still not available. A few efforts have been made to automate the process of constructing or extending existing semantic resources. The most rigorous methodology in the state-of-the-art is the one developed by Zhang and El-Gohary (2016) that is fully-automated with high accuracy. One limitation of this algorithm is the reliance on an existing semantic resource; it, therefore, would not be applicable to such a domain like infrastructure that is out of the vocabulary scope. Thus, there is a need for an automated approach that can not only allow for fast development of highway lexicons but also remove the dependence on other existing semantic models."

Discussions section, Page 21, line 508-519, *"This paper proposes an NLP based methodology to assist professionals in extracting roadway terms and their semantic relations from text documents. A key contribution to the body of knowledge is the novel approach with a new algorithm that allows for automated detection of technical terms and their relations without reliance on existing hand-coded dictionaries as used by previous researchers such as Zhang and El-Gohary (2016). The present framework is not to completely eliminate the human involvement, but is expected to become an enabling tool that can help researchers in the domain quickly develop supporting ontologies and other forms of semantic resources for their specific use cases. With respect to facilitating semantic interoperability for the infrastructure sector, thefindings of this study would accelerate the process of removing the current bottleneck in extensive machine readable dictionaries which are required for an unambiguous data sharing, integration or exchange."*

Reviewer #1:

1. "There are a number of small grammatical errors in the paper and it would benefit from a final English language check (e.g., use of 'to the same' rather than 'of the same')."

Response: Thank you for your comment. The authors have gone through the manuscript and corrected the grammar and spelling errors.

2. "On page 3 you state that research to address the issue of terminology inconsistency has been very limited, but you also point to major decade long efforts such as bSDD. I think you are more concerned with research into automated approaches, and this should be made explicit here."

Response: Yes, this manuscript focuses on the automated approach. The authors have revised the manuscript to highlight the research purpose with higher clarity, as follows.

Page 3, line 65-75, *"Terminology transparency through digital dictionaries like glossaries, taxonomies, ontologies and data dictionaries is identified as a driver of semantic interoperability (Ouksel and Sheth 1999). Although a plethora of semantic*

resources have been introduced to the highway sector, their coverages of terms are still far inadequate for a large number of disciplines, and processes across the highway project life cycle. This is because of the reliance on a tedious and time-consuming approach which requires developers to manually gather and translate knowledge from domain experts or text documents into a machine-readable format. Thus, there is a need for computer-aided methods to remove this knowledge acquisition bottleneck (Mounce et al. 2010), so that digital dictionaries can be quickly constructed to meet a specific need and to keep up with the growth of terms due to rapid applications of new technologies and knowledge."

3. "On page 4 you present no evidence that the growth of terms is exponential - and it is hard to imagine that this could be the case."

Response: We have toned down the sentence. The revised sentences now reads

".....digital dictionaries can be quickly constructed to meet a specific need and to keep up with the growth of terms due to rapid applications of new technologies and knowledge."

4. "On page 6 you could present more explanation as to why reliance on digital dictionaries is becoming a bottleneck."

Response: The authors have included the following information and reference to explain why reliance on digital dictionaries is becoming a bottleneck.

Page 5, line 126-128, *"However, digital dictionaries are typically hand-crafted; they are therefore not available to many domains (Kolb 2008)."*

5. "On page 7 it would be useful to present some information on the performance of the techniques and system that you mention here."

Response: The authors have added several sentences to further discuss the performance of those methods mentioned in the manuscript, as follows.

Page 7, line 158-164, *"For example, the results from a comparative study conducted by Levy et al. (2015) on the accuracy in various tasks and golden standards reveal that Skip-gram outperforms Glove in every experiment and is the winner in most of the tasks, especially on the WordSim Similarity dataset. Among these tasks, the best precision of Skip-gram is .793, while PPMI and Glove achieve the highest score of .755 and .725 respectively."*

6. "On page 8 isn't it more important whether efforts like bSDD are complete rather than how long they take to create?"

Response: The focus of this paper is to reduce the reliance on human efforts in developing dictionaries such as bSDD. The authors have revised the paragraph with new sentences to better describe the idea, as follows.

Page 7 and 8, line 175-196, *"A popular solution to semantic interoperability is to develop taxonomies, ontologies or other forms of digital dictionaries that can provide machine-readable definitions of domain concepts. A plethora of such semantic resources have been developed for the highway industry. However, conventional methods require significant human efforts on knowledge retrieval, ontology construction and validation. The pioneer in this line of research is the e-COGNOS ontology (Wetherill et al. 2002; Lima et al. 2005) which formulates the execution process of a construction project as an explicitly interactive network of the principal concepts: Actor, Resources, Products, Processes and Technical Topics. The ontology developers of this project reviewed existing taxonomies (BS61000, UniClass, IFC) and construction specific documents, and interacted with the end users to identify relevant concepts and their semantic relations. Industry experts were invited to validate the developed ontology through questionnaires on concept names and relations. Since the introduction of the high-level ontology of e-Cognos, a plenty of ontologies have been built for various aspects of the life cycle of a highway project, for instance, highway construction taxonomy (El-Diraby and Kashif 2005; El-Diraby et al. 2005), freight ontology (Seedah et al. 2015a), and the ontology of urban infrastructure products (Osman and Ei-Diraby 2006). Like the e-Cognos project, these studies also relied on domain experts for the construction of their semantic products. The limitation regarding time and labor costs of the ad-hoc traditional methodology has created a bottleneck to the progress in enabling semantic interoperability. In addition, the existing ontologies primarily focus on the description of concepts, and the heterogeneity of concept names is usually neglected. Therefore, research is needed not only to automate the process of formulating domain concepts but also to incorporate term heterogeneity into ontologies."*

7. "On page 9 how do we know that the coverage of this corpora is sufficient and complete? How about terms from corpora in other English language speaking countries? Or corpora from countries which use languages other than English? Need to scope your work better here."

Response: Thank you for your comment. The present framework is limited to American-English text documents. The authors have revised the associated sentence regarding this issue to clearly define the scope of the work, as follows.

Page 10, line 261-263*, "The goal of this research is to propose an NLP-based methodology that can automate the process of extracting roadway technical terms and their semantic relations from American-English roadway documents."*

8. "On page 10 you should discuss the implication of removing tables and equations from your corpus."

Response: The authors have included the following sentences to discuss the implication of removing tables and equations.

***Page 11, line 277-280**, "The removal of these features may slightly reduce the corpus size, and accordingly affect the training dataset; however, it is necessary since words in tables and equations are not organized in the formal structure of a sentence and therefore the NLP algorithm may extract unreal noun phrases."*

9. "On page 14 you should explain what a 'one-hot vector' is."

Response: The authors have included the following sentences to explain the 'one-hot' vector.

***Page 16, line 405-408**, "In this model, a word in the corpus vocabulary is encoded as a "one-hot" vector which is a vector in which only one element at the index of the word in the vocabulary is set one, and all other items are zero. For example, the one-hot vector of k^{th} word in the vocabulary with the size of V will be $\{x_1=0, x_2=0, ..., x_k=1, ..., x_V=0\}$."*

10. "In the discussion and conclusions there is no consideration of what is good enough in a lexicon. is your 81% precision sufficient? Is a F-measure of 65% sufficient?"

Response: The authors understand that the current performance is not high enough for a fully automated process, when 35 percent of the tested terms are out of the vocabulary and 20% of the answers are incorrect. Therefore, the authors have suggested that the most immediate value of this system is significant reduction of human efforts rather than completely removing the human involvement. Users can use this proposed framework as an assistant tool and the automated result may need a manual review. The following statements have been included in the discussions and conclusions sections of the manuscript to discuss the above issues.

***Discussions section, page 21, line 511-516,** "The present framework is not to completely eliminate human involvement, but is expected to become an enabling tool that can help researchers in the domain quickly develop supporting ontologies and other forms of semantic resources for their specific use cases."*

***Conclusions section, page 23, line 569-573,** "Although a significant improvement has been made in comparison with an existing thesaurus database, 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."*

11. "It would be useful to have the final lexicon available online somewhere so that readers of the paper can access and assess the results of this work."

Response: The authors have included a link where the model and datasets from this study can be found at the end of the introduction section, as follows.

Page 4, line 94-95. *"A Java package and several datasets resulting from the study can be found at https://github.com/tuyenbk/mvdgenerator."*

Reviewer #2:

1. "The authors might want to revise the manuscript title to better reflect the scope of research. Particularly, the text corpus came from only roadway design guidelines and does not constitute the entirety of civil infrastructure. While the manuscript places an equal emphasis on the approach as well as on the research product (i.e. InfraLex), the manuscript title should be more carefully crafted into something more accurate"

Response: Thank you for your comment. The authors have revised the title to better reflect the research focus which is an automated approach, and the scope within the roadway sector as follows "NLP-based approach to classifying heterogeneous terms for unambiguous exchange of roadway data."

2. "Several statements within INTRODUCTION need to be supported.
2.1 Page 2, Lines 38-40, "The major cost was time spent...a useful format."
2.2 Page 2, Lines 48-49, "Polysemy and synonymy are two ...data sources."
2.3 Page 3, Lines 61-63, "However, research to address...very limited.""

Response: The authors have included references to support the above arguments. The following are the revised sentences.

Page 2, Line 36-40, *"The inadequate interoperability cost is estimated to be over $15.8 billion per year in the U.S. capital facilities industry as reported by the National Institute of Standard and Technology (NIST); and the largest cost item is the laborious work for finding, verifying, and transferring facility and project information into a useful format during the operation and maintenance stage (Gallaher et al. 2004)."*

Page 3, Line 52-54*, "Polysemy and synonymy are two major linguistic obstacles to semantic integration and use of a multitude of data sources (Noy 2004)."*

Regarding the last statement (page 3, lines 61-63), this is based on the literature review conducted by the authors. The following is the revised statement.

Page 3, Line 67-69, *"Although a plethora of semantic resources have been introduced to the highway sector, as shown in the literature review, their coverages of concepts are still inadequate and the inclusion of multiple names to the same concept is still limited."*

3. "The statement citing the cost of interoperability issue on Page 2 (between lines 35 - 38) does not necessarily apply to the target of the paper, civil infrastructure (or more precisely roadway

design). The citation from NIST was about the capital facilities, which typically refer to buildings and industrial facilities and not horizontal construction."

Response: Since the building and roadway sectors share the same interoperability issues, the intention of the authors here is to provide a tangible evidence about the cost of inadequate interoperability in order to highlight the importance of addressing the same issue in the highway sector. The authors have revised the manuscript as follows to explain that idea.

Page 2, Line 34-46, *"The interoperability issue has been widely recognized as a key obstacle blocking the flow of digital data throughout the entire project life cycle. The inadequate interoperability cost is estimated to be over $15.8 billion per year in the U.S. capital facilities industry as reported by the National Institute of Standard and Technology (NIST); and the largest cost item is the laborious work for finding, verifying, and transferring facility and project information into a useful format during the operation and maintenance stage (Gallaher et al. 2004). 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 major domains in the construction industry, has not yet successfully facilitated a high degree of interoperability (Lefler 2014); huge cost savings would be achieved if roadway data is seamlessly shared across project phases and among state and local agencies.*

4. "The authors seemed to suggest that very few research looked into the terminology inconsistency issues in construction whereas a recent paper published in JCCE attempted to address this particular issue in the transportation sector. The paper is entitled "Ontology for Querying Heterogeneous Data Sources in Freight Transportation" and the authors are advised to review it in order to highlight their research uniqueness and contribution."

Response: Thank you for your recommendation. The authors have reviewed this recommended article and the following discussion on the gap addressed in that publication and others has been added to the revised manuscript.

Page 8, Line 186-196, *"Since the introduction of the high-level ontology of e-Cognos, a plenty of ontologies have been built for various aspects of the life cycle of a highway project, for instance, highway construction taxonomy (El-Diraby and Kashif 2005; El-Diraby et al. 2005), freight ontology (Seedah et al. 2015a), and the ontology of urban infrastructure products (Osman and Ei-Diraby 2006). Like the e-Cognos project, these studies also relied on domain experts for the construction of their semantic products. The limitation regarding time and labor costs of the ad-hoc traditional methodology has created a bottleneck to the progress in enabling semantic interoperability. In addition, the existing ontologies primarily focus on the description of concepts, the heterogeneity of concept names is usually neglected. Therefore, research is needed not only to automate the*

process of formulating domain concepts but also to incorporate term heterogeneity into ontologies."

5. "The authors used a single source of citation from arXiv to support their claim that Skip-Gram model outperforms other methods like LSA and so was adopted in the reported research. This is a critical research design decision and should be more carefully evaluated (and supported). arXiv does not require peer reviews and its publications are only moderated to check against obvious policy violations."

Response: Thank you for your comment. The authors have added other evidences from several peer-review publications to support the selection of this research tool. Below is the revised discussion explaining the tool selection.

Page 6-7, Line 154-164, *"There are contradictive recommendations on the wining model in the literature. The authors of Glove suggested that their model out-performs over Skip-Gram and others in the state of the art. However, a number of independent benchmarking experiments have consistently indicated the outperformance of the Skip-gram model to its alternatives. For example, a comparative study conducted by Levy et al. (2015) on the accuracy in various tasks and golden standards reveals that Skip-gram outperforms Glove in every experiment and is the winner in most of the tasks, especially on the WordSim Similarity dataset. Among these tasks, the best precision of Skip-gram is .793, while PPMI and Glove achieve the highest score of .755 and .725 respectively. The out-performance of Mikolov's model on the similarity task is confirmed in another benchmarking study (Hill et al. 2015) where this model is also found as the winner in most of the tests."*

6. "How do the results from the step 'Noun phrase detection' look like? Porter stemming (or any stemming algorithm) is known to over or under stem terms at times. How was over or under stemming managed when the authors were generating their bag of noun phrases? Why wouldn't the suffix in the word "undivided" in Table 6 be removed after the authors employed the Porter stemming algorithm?"

Response: The authors indeed implemented the Pling Stemmer rather than Porter, but the initial manuscript did not update accordingly." Thank you for capturing this inconsistency. Since the Pling stemmer specifically stems English plural nouns to its singular form, other POS such as adjectives will not be affected. This is the reason why the word "undivided" is not affected. In regards to the errors of over and under-stemming, there are evidences from the literature review that suggest that even though there are still errors in the existing stemmers, they are good enough to not have negative effects on the overall performance of NLP applications. Below is the revised discussion on the stemming process in the proposed method.

Page 12, Line 311-319, *"Stemming is a popular process to reduce words to their stems. Despite the fact that none of the existing algorithms can completely eliminate the errors of over and under stemming, they are good enough to not degrade the overall performance*

of NLP application (Jivani et al. 2011). This study implements the Pling stemmer (Suchanek et al. 2006), which stems an English noun to its singular form, to normalize plural nouns in the corpus. One advantage of this algorithm is the utilization of both syntactic rules and the vocabulary in a dictionary; hence the mis- or over-stemming errors that take off a true suffix can be reduced."

7. "A manual evaluation process was still employed to remove inadequate or meaningless terms from the term candidate list. Could this process become the road block when the authors need to scale up their research?"

Response: The authors understand that the proposed framework is to reduce human efforts in developing semantic resources rather than completely replace domain experts. Manually reviewing of terms would still be a much easier and less time-consuming task than reviewing a plethora of domain text documents. In addition, the authors suggest a method to reduce both the laborious work and the number of real terms automatically removed (the paragraph discussing this method is included below). Using this method, a domain expert does not need to review the entire list. In our experiment, it took a graduate student only 8 hours to finish his task. However, further research is needed to enhance the accuracy of term extraction, so that human involvement would continue to be reduced.

Page 14-15, Line 362-376, *"To automatically remove candidates that are unlikely to be real terms, a threshold C-value can be used. However, doing this may eliminate the real terms that appear in the bottom due to their low frequencies. Manual evaluation of the entire candidate list would avoid the removal of real terms with low C-values. To minimize both laborious work and the number of true terms wrongly discarded, the authors suggest the following method to identify the threshold value. The ranked list of candidates is divided into groups of around 200 items. A graduate student with a civil engineering background was asked to utilize a bottom-up approach to evaluate group by group and stop at which the percentage of actual terms achieved 80 percent. Users can choose a higher percentage limit in cases where the accuracy is critical. Table 2 illustrates the evaluation results for several excerpts of the extracted term candidates. The precision values, which represent the percentages of real terms in these groups, are presented in Figure 3. As shown in the figure, precision values are less than 80 percent for groups with c-values less than 50. This value is set as the threshold for the acceptance of term candidates. The final selected list is comprised of nearly 6,000 multi-word roadway technical terms."*

8. "How were the precision rates in Figure 3 determined? These rates seemed to provide critical information for the manual evaluation upon the term candidates and it is unclear how these precision rates were obtained prior to the manual evaluation."

Response: The precision rate is the percentage of real terms in a group of automatically extracted terms. The conclusion of "real" or "unreal" is based on the manual evaluation process. Below is the revised discussion on the evaluation method and how the precision rates are determined.

Page 15, Line 367-373, *"The ranked list of candidates is divided into groups of around 200 items. A graduate student with a civil engineering background was asked to utilize a bottom-up approach to evaluate group by group and stop at which the percentage of actual terms achieved 80 percent. Users can choose a higher percentage limit in cases where the accuracy is critical. Table 2 illustrates the evaluation results for several excerpts of the extracted term candidates. The precision values, which represent the percentages of real terms in these groups, are presented in Figure 3."*

TABLE 2: Excerpts of the extracted candidate terms

Term candidate	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

9. "Results in Table 7 are not sensitive to the different parameter settings. Table 7 can be removed from the manuscript."

Response: Thank you. The authors have also recognized that there is no consistent relation pattern between the performance and the configuration of these parameters. However, the authors would like to respectfully keep this table in the manuscript to visualize that finding.

TABLE 7: Performance of the synonym matching task with various training settings

Parameter changed	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

.

10. "The body of knowledge does not lie within the method and was more on the research product. This is a relatively weak contribution"

Response: The authors have revised the literature review section to focus more on the state-of-the-art approaches and discussions on the contribution of this manuscript to the body of knowledge. Below is the revised body of knowledge section.

Page 7-10, Line 175-258,

"Related studies

A popular solution to semantic interoperability is to develop taxonomies, ontologies or other forms of digital dictionaries that can provide machine-readable definitions of domain concepts. A plethora of such semantic resources have been developed for the highway industry. However, conventional methods require significant human efforts on knowledge retrieval, ontology construction and validation. The pioneer in this line of research is the e-COGNOS ontology (Wetherill et al. 2002; Lima et al. 2005) which formulates the execution process of a construction project as an explicitly interactive network of the principal concepts: Actor, Resources, Products, Processes and Technical Topics. The ontology developers of this project reviewed existing taxonomies (BS61000, UniClass, IFC) and construction specific documents, and interacted with the end users to identify relevant concepts and their semantic relations. Industry experts were invited to validate the developed ontology through questionnaires on concept names and relations. Since the introduction of the high-level ontology of e-Cognos, a plenty of ontologies have been built for various aspects of the life cycle of a highway project, for instance, highway construction taxonomy (El-Diraby and Kashif 2005; El-Diraby et al. 2005), freight ontology (Seedah et al. 2015a), 190 and the ontology of urban infrastructure products (Osman and Ei-Diraby 2006). Like the e-Cognos project, these studies also relied on domain experts for the construction of their semantic products. The limitation regarding time and labor costs of the ad-hoc traditional methodology has created a bottleneck to the progress in enabling semantic interoperability. In addition, the existing ontologies primarily focus on the description of concepts, the heterogeneity of concept names is usually neglected. Therefore, research is needed not only to automate the process of formulating domain concepts but also to incorporate term heterogeneity into ontologies.

Another strategy for semantic interoperability targets at the heterogeneity of concept names rather the concept description as in an ontology model. A few frameworks to assist practitioners in precisely mapping data labels from heterogeneous sources have been introduced for various construction sectors. In the building sector, buildingSMART proposed a novel framework, namely IFD (International Framework for Dictionaries) (ISO 12006-3) for developing a multilingual data schema in which each concept can have multiple names in different languages. With IFD, the identity of a concept is defined by a Global Unique ID (GUID) rather than its name; hence an IFD-based data exchange mechanism is able to eliminate the semantic mismatches due to the name inconsistency (IFD Library Group ; Hezik 2008).. The buildingSMART data dictionary (bSDD) (buildingSMART 2016) is the first digital library of building concepts that is crafted in the IFD structure. Each concept in bSDD consists a set of synonymy names not only in English but also in computer-coded languages (e.g., IFC-Industry Foundation Classes) and in other human languages (e.g., French, Norwegian). Therefore, a complete bSDD would enable digital data in regardless of languages to be sharable and unambiguously reusable. Yet, its size remains limited as the identification of these sets of synonyms is labor and time

extensive. In the transportation sector, there has been a shortage of research efforts targeting the heterogeneity of data names at the database level until recently. Seedah et al. (2015b) proposed a role-based classification schema (RBCS) to classify data in freight databases. RBCS defines nine distinct groups of roles that are time (year, month), place (city name, population), commodity (liquid, value), link (roadway name, width), mode (truck, rail), industry (company name, sales), event (accident, number of fatalities), and human (officer, driver age). The authors argue that once the data elements across separate databases are categorized using this standard system, it becomes easier for practitioners to identify the semantic relatedness in their definitions. However, even if RBCS is successfully applied to all freight databases, identifying the exact type of relation (synonym, functional relation) between two data elements in the same category is still a challenging task.

In attempts to reduce laborious work on defining concepts, a few researchers have sought to propose semi-automated and automated methods for identifying semantic relations among technical terms. Abuzir and Abuzir (2002) developed the ThesWB system which utilizes hand-coded syntax patterns to detect lexical relations between civil engineering terms from HTML web pages. The performance of ThesWB was not reported, but it is not likely to be high since rule-based approaches are repeatedly criticized for not being able to capture all the variant ways to present relations among terms in natural language (Marcus 1995; Navigli and Velardi 2010). Rezgui (2007) suggested a more sophisticated approach that is based the statistics of word occurrence rather than predefined rules to extract potential pairs of related terms from domain text documents. This method implements TF-IDF to evaluate the importance degree of a keyword to the examined domain; and analyzes the co-occurrence frequencies using Metric Clusters to assess the potentiality that exists a semantic relation within a given pair of important keywords. These potential relationships are then validated and categorized by domain experts. Since only pairs of terms that occur in the same sentence are considered, equivalent terms which are used interchangeably could not be captured. In another study to identify semantic relations, Zhang and El-Gohary (2016) proposed a fully automated methodology for both tasks of retrieving related candidate and classifying the relations. This algorithm was reported to achieve an average precision of nearly 90 percent in the relation classification task. However, the algorithm identifies potentially related concepts based on the pre-defined lexical relations provided in WordNet, a generic lexicon that lacks concepts in many construction sectors including the civil infrastructure, it would not be scalable well on matching terms in these domains.

As shown in the literature review, there are numerous research efforts in developing ontologies for the highway sector. However, the existing ontologies are mainly hand-coded through manual processes of knowledge acquisition and formally describing them in a digital format. This ad-hoc approach has created a bottleneck in facilitating the semantic interoperability for the whole industry and as a result, semantic resources for many aspects of a project are still not available. A few efforts have been made to automate the process

of constructing or extending existing semantic resources. The most rigorous methodology in the state-of-the-art is the one developed by Zhang and El-Gohary (2016) that is fully-automated with high accuracy. One limitation of this algorithm is the reliance on an existing semantic resource; it, therefore, would not be applicable to such a domain like infrastructure that is out of the vocabulary scope. Thus, there is a need for an automated approach that can not only allow for fast development of highway lexicons but also remove the dependence on other existing semantic models."