# NLP-based approach to classify 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 to support decision making in asset management. The issue of semantic heterogeneity may lead to the lack of common understanding to the same data between the sender and receiver. Semantic resources, lexicons, and ontologies, that formally describe the definitions of data labels will enable their meanings to be precisely understood by computer systems. However, the current manual process of developing these dictionaries for the civil infrastructure sector is laborious and time-consuming due to the lack of an effective automated method. This paper presents a novel methodology to construct an automatically-generated lexicon, namely RoadLex, that organizes roadway technical terms in a lexical network manner. Natural Language Processing (NLP) techniques and the C-value method are first implemented to extract commonly used technical terms from a corpus of roadway design guidelines collected from across the State Departments of Transportation. A model for measuring the semantic similarity is then trained on the data of context words of these terms in the corpus using the Skip-gram neural network model. This semantic model is then utilized by a proposed term classification algorithm that measures the semantic similarity between terms and assigns relation types (synonyms, hyponyms, and functional relations) to each pair of related terms. The final network of terms is organized in a Wordnet-like format in which terms are

[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.

grouped into sets of synonyms, and these groups are connected to one another through the hyponym and functional relations. The proposed methodology was evaluated by conducting an experiment comparing the automatically-identified synonyms in RoadLex with a human-constructed golden standard dataset obtained from Wikipedia. The result shows that the proposed model achieved a precision of over 80 percent.

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

## INTRODUCTION

The implementation of advanced computerized technologies such as 3D modeling and Geographic Information System (GIS) throughout the life cycle of a civil infrastructure project has allowed a large portion of project data to be available in digital format. The efficiency improvement in sharing these data between project participants and stages, will in turn, translate into increased productivity, efficiency in project delivery and accountability. The interoperability issue has been widely recognized as the key obstacle blocking the flow of digital data through the entire project life cycle. The inadequate interoperability is estimated of over $15.8 billion per year in the U.S. capital facility 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 key 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 through the project life cycle and among state and local agencies.

Semantic interoperability, which relates to the issue whereby two computer systems may not have the same understanding to the same piece of data, is a radical barrier to computer-

to-computer data exchange. Due to the fragmented nature of the infrastructure domain, data representation/terminology 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 materials or functions. Walton et al. (2015) suggests the following three reasons for the semantic heterogeneity: (1) isolation in definitions among separate sources, (2) temporary of definitions and (3) variety of data collection methods. Synonymy, in contrast, is associated with the heterogeneity of terms used to represent the same concept. For instance, the longitudinal centerline of a roadway has various representative terms including 'profile', 'crest', 'grade-line' and 'vertical alignment'. Simply mapping of data names will likely lead to the failure of data extraction, or use of wrong data. Thus, addressing the terminology ambiguity due to the semantic inconsistency issue becomes crucial to ensure the common understanding to the same dataset by all software applications and guarantee the extraction of right data and 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). Unfortunately, although, a plethora of semantic resources have been introduced for the highway sector, their coverages of terms are still far inadequate for a large number of disciplines, and processes across the project life cycle. This is because of the reliance on a tedious and time-consuming approach which requires the developers to manually gather and translate knowledge from domain experts or text documents into a machine-readable format. There is a need for computer-aided methods to remove this knowledge acquisition bottleneck (Mounce et al. 2010), such that digital dictionaries can be quickly constructed to meet a specific domain and to keep up with the sustainable growth of new terms arisen

along with new knowledge and technologies.

Recent achievements in accuracy and processing time of advanced Natural Language Processing (NLP) techniques 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 great 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 fast translation of domain knowledge into a computer-readable format which is required for the machine-to-machine based data exchange.

This paper presents an automated approach using NLP to translate text-based domain knowledge into an extensive American-English lexicon of roadway data terminology, namely RoadLex. The lexicon formally organizes technical terms in a lexical hierarchy manner that can serve as the core dictionary in a data integration system. In order to achieve that goal, Natural Language Processing (NLP) techniques and the C-value method (Frantzi et al. 2000) are used to detect technical terms from a corpus of roadway design guidelines collected from across the State Departments of Transportation. A model for measuring the semantic similarity is then trained on the data of context words of these terms in the corpus using the Skip-gram neural network model (Mikolov et al. 2013). This semantic model is then utilized by a proposed term classification algorithm that measures the semantic similarity between terms and assigns relation types (synonyms, hyponyms, and functional relations) to each pair of related terms. The final network of terms is organized in a Wordnet-like format in which terms are grouped into sets of synonyms, and these groups are connected to one another through the hyponym and functional relations. A Java package and a lexicon dataset result 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 like adjective, noun, verb, etc. to each token of a sentence) (Toutanova et al. 2003; Cunningham et al. 2002), and Dependency parser (relationships between linguistic units) (Chen and Manning 2014), to the semantic level like word sense disambiguation (Lesk 1986; Yarowsky 1995; Navigli 2009). NLP methods can be classified into two main groups: (1) rule-based and (2) machine-learning (ML) based methods. Rule-based methods, which rely solely on hand-code syntax rules, are not able to fully cover all human rules (Marcus 1995); and their performance is, therefore, relatively low. In contrast, the ML-based approach is independent of languages and linguistic grammars (Costa-Jussa et al. 2012) as linguistics patterns can be fast learned from even un-annotated training examples. Thanks to its impressive outperformance, NLP research is shifting to statistical ML-based methods (Cambria and White 2014).

## Vector Representation of Word Semantics

Measuring of semantic similarity, which is one of the main NLP-related research topics, aims to determine how much two linguistic units (e.g., words, phrases, sentences, concepts) are semantically alike. For example, a *bike* might be more similar to a *car* than to *gasoline*. The state-of-the-art methodology for 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 that consists of terms organized in a lexical hierarchy of semantic relations such as synonym, attribute, hypernym/hyponym, etc. Computational platforms (e.g., information retrieval) built upon such dictionaries are

able to fast measure the semantic similarity by computing the distances between words in the hierarchy. Hence, this method would be an ideal solution when digital dictionaries are available. However, digital dictionaries are typically hand-crafted; they are therefore not available to many domains (Kolb 2008). The latter method, on the other hand, assess 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 is represented by the Euclidean distance between the corresponding points (Erk 2012). The conventional method to construct the representation vectors of semantic units 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 measure the co-occurrence probabilities. A more advanced approach uses machine learning to train the vector representation of terms. One example of this line of methodology is the Skip-gram neural network model (Mikolov et al. 2013) which predicts the context words of a given input word. The training objective is to minimize the 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 local context data within a context window, Glove trains on the global co-occurrence statistics. There are contradict claims on the best model when the authors of these two learning model both claimed the out-performance of their models to the state of the art. However, a number of independent benchmarking experiments have consistently indicated the outperformance of the Skip-gram model to other alternatives. For example, the results from the study conducted by Levy et al. (2015) on the accuracy of the state-of-the-art semantic models 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 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) when this model is also the winner in most of tests.

The VSM approach has been progressively implemented in the 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 increasingly number of successful use cases in the construction industry has evidently demonstrated that the VSM method can successfully identify the semantic similarity between data labels which is critical to tackle the issue of semantic interoperability in sharing digital data across the life cycle of a highway project.

**Approaches to semantic interoperability in construction**

A popular solution to semantic interoperability is to develop formal semantic digital resources, taxonomies, and ontologies which can enable the definitions of domain concepts to be understandable to computer systems. A plethora of ontologies, taxonomies have been developed. However, the existing products mostly result from the manual processes of knowledge retrieval, and ontology definition 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 implicitly 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 relations among them. Industry experts were invited to validate the developed ontology through questionnaires on terms used an their 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 infrastructure projects, 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 relied on domain experts to define entities and their relations. The limitation regarding time and labor costs of the ad-hoc traditional methodology has created a bottleneck to the progress of facilitating semantic interoperability. In addition, since the primary goal of an ontology is to describe the definitions 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 the ontology.

Another strategy on semantic interoperability focuses on the heterogeneity of concept names rather the concept description in an ontology. A few frameworks to assist practitioners in integrating data 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-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 synonym identification is hardly dependent on a manual process. 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 relation type (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 relation 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 likely to be relative low since rule-based approaches are repeatedly criticized for not being able 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 civil infrastructure, it would not be scalable well on matching terms in these domains.

As shown in the literature review, there is a numerous research efforts in developing ontologies for the the highway sector. However, the existing ontologies for the highway sector are mainly hand-coded through an manual process of knowledge acquisition and formally describing them in digital format. This ad-hoc approach has created a bottleneck to facilitate the semantic interoperability level for the whole industry when semantic resources for many aspects of the 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 an need for an automated approach that can not only allow for fast development of highway lexicon but also remove the dependence on other existing semantic models.

## PROPOSED METHODOLOGY TO AUTOMATED CLASSIFICATION OF

## HIGHWAY TERMS

The goal of this research is proposed an NLP-based methodology to automate the process of knowledge gathering from text and constructing an American-English lexicon of roadway terms named RoadLex. 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 text corpus to develop a Roadway Vector Space Model (Rd-VSM) that presents the semantics of highway terms, and (3) develop an algorithm integrating Rd-VSM and various linguistic patterns to classify relations among technical terms (synonym, hyponym and functional association) which are then used to construct the RoadLex network of roadway terms. Specifically, the procedure followed to compile the RoadLex dictionary is comprised of the following steps: (a) collecting roadway technical documents to compose a domain corpus; (b) extracting multi-word terms from the roadway corpus; (c) prepare the training dataset for training the Rd-VSM model; (d) specifying the parameters of the machine learning model and perform the training task to develop the Rd-VSM model; and (e) designing an algorithm to classify related terms into groups of lexical relations. The below sections discuss these steps in detail.

### Text corpus collection

In order to develop a domain text corpus of the highway sector, the authors collected a plethora of highway engineering manuals and guidelines from the Federal Department of Transportation (DOT) and from 22 State DOTs. 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 structure 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 would 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. The final outcome of this phase is a plain text corpus consisting of 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**

A technical term can be a noun (e.g., roadway, lane, etc.) or be a noun phrase composed of multiple words (e.g., right of way, at grade intersection, etc.). For multi-word terms, what they mean may not be directly interpreted from the meanings of their single words. In order for the Skip-gram model to learn the semantics of multi-word terms, they must be treated as single-word terms by replacing them with connected blocks of word members. As mentioned, a multiple-word terms must be a noun phrase; therefore, noun-phrases will be good term candidate. To detect multiple-word terms, the corpus is first scanned to look for noun phrases and their occurrence frequencies data. This statistical data is then analyzed to assess the importance of these noun phrases to the sector of highway. The process of extracting multi-word terms is discussed in detail in the following sections.

*Noun phrase extraction*

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

iii Noun phrase detection: Linguists argue that a technical term is either a noun (e.g., road) or a noun phrase (NP) (e.g., right of way) that frequently occurs in domain

text documents (Justeson and Katz 1995). Thus, NPs are good multi-word term candidates. Table 1 presents the proposed extraction patterns which are modified from the filters suggested by Justeson and Katz (1995) to extract NPs. The tagged corpus is thoroughly scanned, and sequences matching to the noun phrase patterns is collected. In addition, in order to avoid discrimination among the syntactic variants of the same term, for example 'roadway' and 'roadways', the collected NPs need to be normalized. The following are two types of syntactic variants and the proposed normalization methods.

- Type 1 - Plural forms, for example 'roadways' and 'roadway'. The Porter stemming algorithm (Porter 1980), which can allow for automated removal of suffixes, is applied on the extracted noun phrases to normalize plural nouns (NNS) into single nouns (NN). Since the stemming algorithm affects only on the NNS token of a Noun phrase, the issue of over and under stemming can be minimized/eliminated.
- Type 2 - Preposition noun phrases, for example 'roadway type' and 'type of roadway'. In order to normalize this type of variant, the form with preposition is converted into the non-preposition form by removing the preposition and reversing the order of the remaining portions. For instance, 'type of roadway' will become 'roadway type'.

The first column in Table 2 represents several examples of the NP list retrieved from this phase. Since a NP is not certainly a technical term, ones that are clearly unlike to be terms should be excluded from the candidate list. A basic indicator is the occurrence frequencies as technical terms would repeatedly occur in the domain text documents. To eliminate unlikely candidates, a threshold of frequency can be applied. If the user 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 is to extract 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 a NP is a multi-word term (Frantzi et al. 2000). There are a number of methods proposed for estimating termhood (the degree that a linguistic unit is a domain-technical concept), such as TF-IDF (Sparck Jones 1972; Salton and Buckley 1988), C-Value (Frantzi et al. 2000), Termex (Sclano and Velardi 2007). These methods are based on the occurrence frequencies of NPs in the corpus. Among these methods, Termex outperformed other methods on the Wikipedia corpus, and C-Value was the best on the GENIA medical corpus (Zhang et al. 2008). This result indicates that the C-value method is more suitable for term extraction from a domain corpus rather than a generic corpus. For this reason, the C-value has been widely used to extract domain terms in the biomedical field, for instance studies performed by Ananiadou et al. (2000), Lossio-Ventura et al. (2013), and Nenadić et al. (2002). Since the corpus used in this study was mainly collected from technical domain documents, C-value would be the most suitable for the termhood determination task. The C-value measure, as formulated in Equation 1, suggests that the longer a NP is, the more likely that is a term; and the more frequently it appears in the domain corpus, the more likely it will be a domain term.

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

Where:

**a** is a candidate noun phrase

$|\mathbf{a}|$ is the length of noun phrase $a$

**f** is the frequency of $a$ in the corpus

**Ta** is the set of extracted noun phrases that contain $a$

**P(Ta)** is the size of Ta set.

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

To remove candidates that are unlikely to be real terms, a threshold C-value can be used or the entire candidate list should be manually evaluated by industry experts. Manual evaluation would avoid the removal of real terms with low C-values. To minimize both laborious work and the number of true terms wrongly discarded, a threshold can be applied. The authors suggest the following method to identify this limit value. The ranked list of candidate were first divided into groups of 100 items. A graduate student with civil engineering background was asked to utilize a bottom-up approach to evaluate group by group and stop at which the percentage of real-term achieved 80 percent. Users can choose a higher percentage in cases that extracting of rare terms 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 70. This value is set as the threshold for the acceptance of term candidates.

### Construction of term space model

This step aims at processing the collected text corpus and collecting the training data for developing the Rd-VSM model. Skip-gram (Mikolov et al. 2013), which is an un-supervised machine model, was employed to learn the semantic similarity among words in the text corpus. The Skip-Gram model requires a set of training data in which the input data is a linguistic unit (word or term), and the output data is a set of context words which are closed to the input unit in the corpus. In order to collect this training dataset, the unannotated

highway corpus is scanned to capture instances of terms and their corresponding context words. Each occurrence of a word will correspondingly generate a data point in the training dataset.

Before collecting the training dataset, an additional step is needed to handle the issue related to multi-word terms. Since document scanning is on a word-by-word basis, the corpus must be adjusted so that multi-word terms can be treated like single words. To fulfill that requirement, every occurrence of a certain multi-word term in the corpus is replaced with a single unit that is compiled by connecting all the individual words. For instance, 'vertical alignment' becomes 'vertical-alignment'.

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

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

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

W matrix which is the output of the learning process, is a representation vector of a word in the corpus vocabulary. The similarity among these vectors represent the similarity of the context of the corresponding words. The bullets below explains how the context vector of an input word is predicted using the the parameter matrices resulted from the learning process. As we can see, the similarity between two predicted context vectors depends only the similarity between the input representation vectors; thus, these vectors are used to represent the semantics of words.

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

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

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

The model includes three major parameters that are *frequency threshold*, *hidden layer size* and *window size* (see Table 3). To eliminate those data points with low frequencies of occurrence that are unlikely to be technical terms, Word2vec allows for the use of *frequency threshold*. Any word with the rate lower than the limit will be ignored. Radim (2014) suggests a range of (0-100) depending on the data set size. Setting this parameter high will enhance the accuracy, but many true technical terms would be out of vocabulary. A preliminary study based on the preliminary corpus with only several millions of words shows that with the frequency of 20, there are very few non-technical terms involved in the training dataset. Hence, with the larger dataset to be collected, this parameter can be higher and up to around 50. The second important parameter is *layer size* which determines the number of nodes in the hidden layer. This parameter highly affects the training accuracy and processing time. A larger layer size is better in terms of accuracy, but this will be paid off by the running time. The reasonable values for this parameter are from ten to hundreds (Radim 2014).

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

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

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

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

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

Where: n is the hidden layer size.

**Construction of term lexical hierarchy**

The purpose of this module is to construct RoadLex, a lexicon of civil engineering technical terms. A lexicon, also known as a lightweight knowledge base, typically includes terms and relations. The core relations of a lexicon are synonym (meaning equivalence), hypernym-hyponym (also known as IS-A or parent-child relation), attribute (concept property), and association (e.g. part-of) (Jiang and Conrath 1997; Lee et al. 2013). Two terms that relate

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. Table 4 illustrates a list of nearest terms of 'roadway'. In this list, the true synonyms are 'highway' (1), 'traveled-way' (2) and 'road' (4); the attributes include 'roadway-section' (3), 'roadway-shoulder' (12); and 'adjacent-roadway' (7) and 'undivided' (37) are hyponyms which show different types of roadway.

The specific objective of this task is to detect the semantic relations among terms which are used for rearranging the nearest terms obtained from the Rd-VSM model. Algorithm 1 shows the design pseudo code for classifying the nearest terms of a given target term. The algorithm utilizes linguistic rules and clustering analysis to organize the nearest list into the following three groups: (1) attribute, (2) hyponym, and (3) synonym/sibling. The algorithm detects terms belonging to the first two categories using linguistic patterns; and employ 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 *Noun2* concept.

*Synonyms*

The remained nearest words will fall into the third group. However, some of them have far or even no relation with the target word. In order to address this issue, this research employed the K-mean clustering algorithm (MacQueen 1967) to split the remained list into three distinct layers based on the similarity score. The terms in the last group are unlikely to be a synonym or sibling; and thus, are removed from the nearest list. The output of the proposed algorithm is a list of classified nearest terms. Table 6 shows one example for the output retrieved from the algorithm.

**Algorithm 1** Semantic relation classification algorithm

**Inputs**: term $t$, list of nearest terms $N$, full list of 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

## PERFORMANCE EVALUATION

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

overall performance of a system.

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

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

$$F - measure = \frac{2.Precision.Recall}{Precision + Recall} \quad (6)$$

Table 7 shows the performance with various training model settings. The parameters of the baseline model are 50, 100 and 5 respectively for frequency threshold, hidden layer and window size. The authors changed these parameters one by one and kept the other ones unchanged to evaluate their effects to the model performance. As presented in the table, the 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 change of other parameters did not improve the performance. Especially, the increase of frequency threshold value has negative impact. This result confirms the reasonable selection of the frequency threshold to eliminate unlikely term candidate in the NP extraction phase.

The proposed model was also compared with the generic Wordnet database. Table 8 presents the comparison of performance between RoadLex (with the 50-100-10 setting) and Wordnet. As shown, RoadLex outperforms Wordnet in all measures, and the combined F-measure is significantly improved (65% compared to 52%). The biggest contribution to the improvement of the overall F-measure is the recall value which represents a better coverage of domain vocabulary of RoadLex.

## 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 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 proposed method has been applied on a roadway corpus of over 16 million words to generate a lexical dataset of more than 6 thousand terms. The present framework is expected to become a 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 respects to the facilitation of semantic interoperability for the infrastructure sector, the implications 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 lexicon dataset developed in this study is expected to become a fundamental resource for a variety of NLP related studies in the civil infrastructure domains. RoadLex can serve as a machine-readable dictionary of domain technical terms. NLP based platforms can utilize this resource for term sense analysis which is crucial for text mining to extract meaningful information from text documents, information retrieval, or natural language based human-machine interaction. Some specific examples of these potential applications are as follows. First, information retrieval systems can use the semantic relations provided by RoadLex to classify project documents by relevant topics by analyzing the keywords in the documents. Second, questionnaire designers can utilize RoadLex to search for synonyms so that appropriate terms can be selected for specific groups of potential respondents who might be from multiple disciplines or regions. Another application is that the query systems for extracting data from 3D engineered models would be able to find alternative ways to query data when users' keywords do not match any entity in the database. Since users have different ways and keywords to query data, the ability to recognize synonyms and related concepts of a query system would provide flexibility to the end user. Also, the developed RoadLex lexicon would enable the matching data items such as (e.g., cost, productivity, etc.) when integrating data from distinct departments or states to develop a national database. This study is also expected to fundamentally transform the way human interacts with machine

as technical terms which are a basic unit of human language can be precisely understood by computer systems. Instead of using computer languages, the end user can use natural language to communicate with computer systems.

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 would enable more technical terms to be covered in RoadLex. Further work is needed to enhance the performance of RoadLex by enlarging the data training set in both size and the number of disciplines involved throughout the life cycle of a highway project, such as asset management, project programming, construction management. The corpus also needs to cover other types of transportation assets like bridge, tunnel, railway, culvert, etc. Another work that can potentially improve the model performance is to distinguish synonym and sibling which are still in the same group in the RoadLex system. When these two lexical relations are separated, the possibility of recognizing a wrong synonym will be reduced; and consequently, the precision value would be enhanced.

## CONCLUSIONS

Data manipulation from multiple sources is a challenging task in infrastructure management due to the inconsistency of data format and terminology. The contribution of this study is a digital lexicon of highway related technical terms (named RoadLex) which can enable a computer to understand semantic meanings of terms. This research employs advanced NLP techniques to extract technical terms from a highway text corpus which is composed of 16 million words built on a collection of design manuals from 22 State DOTs across the U.S. Machine learning was used to train the semantic similarity between technical terms. An algorithm was designed to classify the nearest terms resulted from the semantic similarity model into distinct groups according to their lexical relationships. This algorithm was employed to develop the RoadLex database.

The developed lexicon has been evaluated by comparing the results obtained from the

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

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

## REFERENCES


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

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

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

buildingSMART (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*.

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). "Roadway safety data interoperability between local and state agencies." *Report no*.

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

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

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

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*

683 *Language Resources and Evaluation (LREC2002)*, 2155–2162.

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

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

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

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

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

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

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

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

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

702 Seedah, D. P., Choubassi, C., and Leite, F. (2015a). "Ontology for querying heteroge-
703 neous data sources in freight transportation." *Journal of Computing in Civil Engineering*,
704 04015069.

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

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

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

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

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

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

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

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

Yalcinkaya, M. and Singh, V. (2015). "Patterns and trends in building information modeling (bim) research: A latent semantic analysis." *Automation in Construction*, 59, 68–80.

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

Zhang, J. and El-Gohary, N. (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 Tables

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 in RoadLex

| **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 RoadLex

| **Lexicon** | **Precision (%)** | **Recall(%)** | **F (%)** |
|---|---|---|---|
| Wordnet | 76 | 40 | 52 |
| **RoadLex** | **81** | **54** | **65** |

## List of Figures

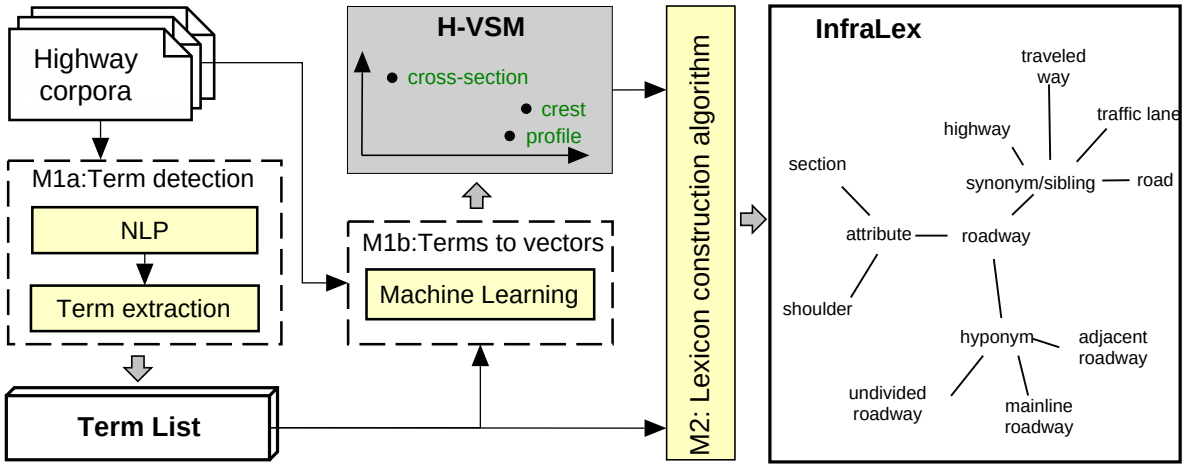


FIG. 1: Overview of the proposed methodology

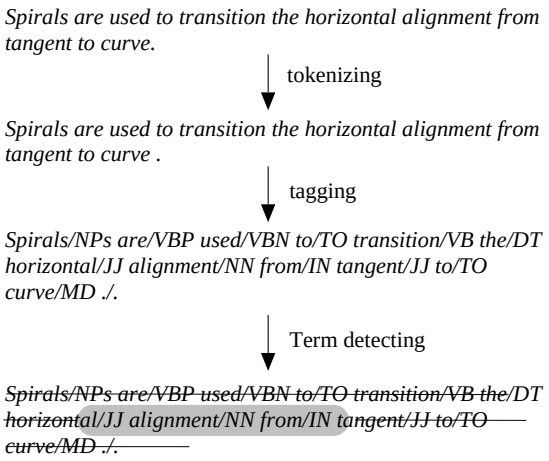


FIG. 2: Linguistic processing procedure to detect technical terms

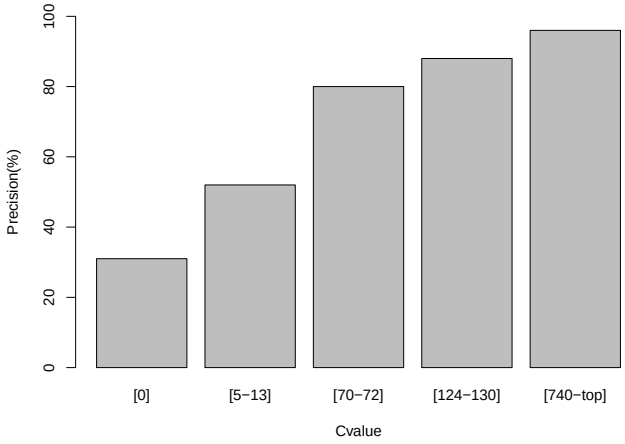


FIG. 3: Multi-word term extraction evaluation

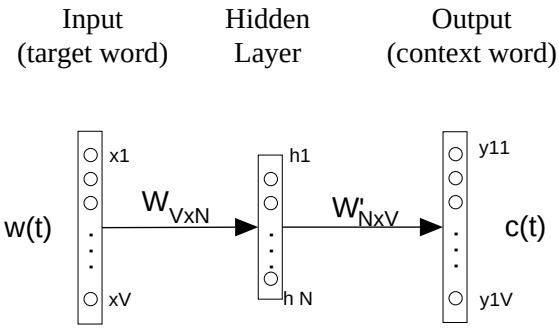


FIG. 4: Skip-gram model

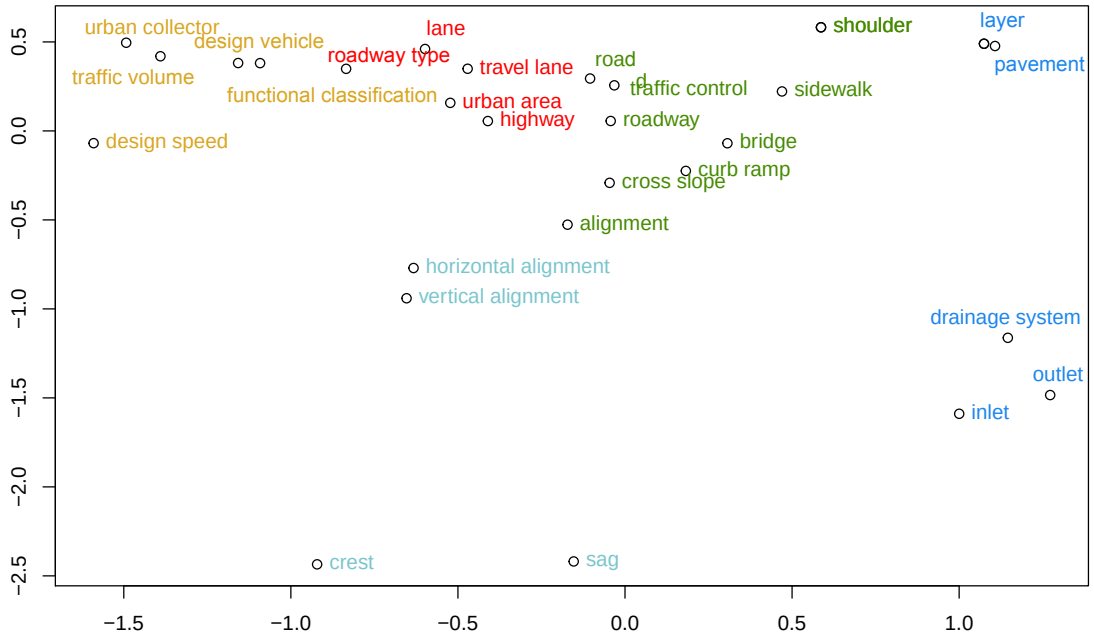


FIG. 5: Highway term space model (Rd-VSM)