**Response to Reviewer Comments**

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

**Manuscript #:** CPENG-1952

The authors wish to thank the editors and reviewers for their time in effort in reviewing our manuscript. We hope the changes listed have made the manuscript 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: Thank you for your comment. The authors have revised the manuscript in according with the recommendations from the reviewers. Please see the responses below to the reviewers’ comments. With respects to highlighting the research contribution, the authors have revised the manuscript with the aim to provide readers with a clearer discussions and justification for the research value. The following are examples of revised paragraphs discussing the contribution of the research in different sections through the 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). 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 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 need and to keep up with the sustainable growth of terms arisen along with new knowledge and technologies.”*

***Related studies section, page 10, line 245-257,*** *“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 level for the whole industry when 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 506-517,*** *“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 interfere, 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 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.”*

**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 correct 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: Thank you for your comment. Yes, this manuscript focuses on the automated approach. The authors have revised the manuscript to better highlight the research purpose, 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). 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 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 need and to keep up with the sustainable growth of terms arisen along with new knowledge and technologies.”*

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: Thank you for your comment. The term “exponential” has been replaced by “sustainable” on page 3, line 74.*

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

*Response: Thank you for your comment. The authors have included the following information to explain why reliance on digital dictionaries is becoming a bottleneck.*

***Page 5, line 125-127,*** *“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: Thank you for your comment. The authors have added several sentences discussing the performance of the mentioned methods to the manuscript, as follows.*

***Page 6-7, line 157-161,*** *“For example, the results from 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.”*

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

*Response: Thank you for your comment. The emphasis of this paper is to reduce the reliance on human efforts in developing dictionaries such as bSDD. The author have revised the paragraph with several new sentences added to better describe the idea, as follows.*

***Page 7 and 8, line 174-195,*** *“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 numerous 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, 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 260-262****, “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: Thank you for your comment. The authors have included the following sentences to discuss the implication of removing tables and equations.*

***Page 11, line 276-279****, “The removal of these features slightly reduces 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.”*

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

*Response: Thank you for your comment. The authors have included the following sentences to explain the ‘one-hot’ vector.*

***Page 16, line 406-406****, “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 kth word in the vocabulary with the size of V will be {x1=0, x2=0, ..., xk=1, ..., xV=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: Thank you for your comment. 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 this system is expected to significantly reduce human efforts rather than completely remove the human interference. 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 510-514,*** *“The present framework is not to completely eliminate the human interfere, 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 567-571,*** *“Although a significant improvement is shown 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: Thank you for your comment. 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 result 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 classify 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: Thank you for your comment. 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 of 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, this is based on the literature review conducted by the authors. The following is the revised statement.*

***Page 3, Line 67-69,*** *“Unfortunately, although, a plethora of semantic resources have been introduced for 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: Thank you for your comment. Since the building and roadway sectors share the same issue of interoperability. The authors intended to provide an evidence about the cost of inadequate interoperability to highlight the importance of addressing that 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 through the entire project life cycle. The inadequate interoperability cost is estimated of 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 through 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 comment. The authors have reviewed this recommended article and the following discussion on the gap in that publication and others has been added to the revised manuscript.*

***Page 8, Line 185-195,*** *“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 discussions reasoning for the tool selection.*

***Page 6-7, Line 153-163,*** *“There are contradict 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, the results from 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: Thank you for your comment. 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 309-317,*** *“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?”

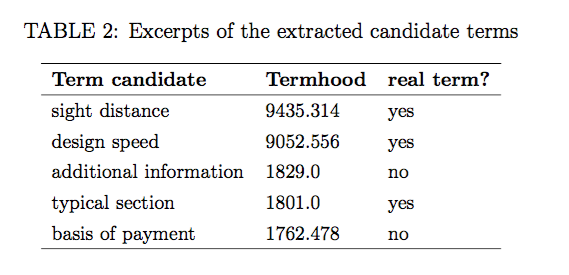
*Response: Thank you for your comment. The authors understand that the proposed framework is to reduce human efforts in developing semantic resources rather than completely replace domain experts. Manually reviewing of thousands of terms still be a much easier and less time-consuming task than reviewing a plethora of domain text documents. In addition, the authors suggest a method to reduce both the laborious work and the number of real term 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 4 hours to finish his task. However, further research is needed to enhance the accuracy of term extraction, so that human interfere would continue to be reduced.*

***Page 14-15, Line 360-374,*** *“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 candidate 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 that 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.”

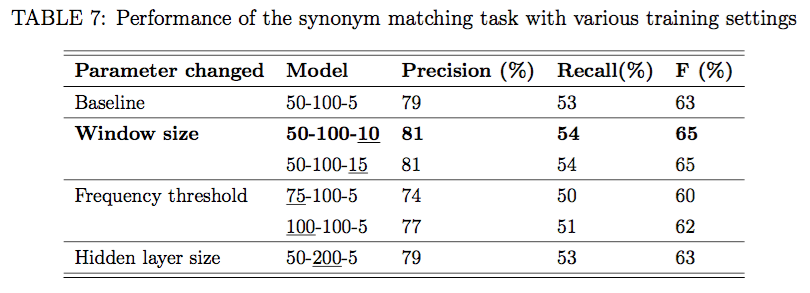
*Response: Thank you for your comment. The precision rate is the percentage of real terms in a group of automatically extracted terms. The conclusion of “real” or “unreal” is based on the manual evaluation process. Below is the revised discussion on the evaluation method and how the precision rates are determined.*

***Page 15, Line 368-375,*** *“The ranked list of candidate 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 that 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.”*

**

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

*Response: Thank you for your comment. The authors have also recognized that there is no consistent relation pattern between the performance and the configuration of these parameter. However, the authors have decided to keep this table in the manuscript to visualize that finding.*

*.*

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: Thank you for your comment. 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 o the body of knowledge. Below is the revised body of knowledge section.*

***Page 7-10, Line 173-257,***

*“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 numerous 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 level for the whole industry when 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.”*