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

**Manuscript #**: CPENG-1952

**Abstract**

The inconsistency of data terminology due to the fragmented nature of the highway industry has imposed big challenges on integrating digital data from distinct sources. The issue of semantic heterogeneity may lead to the lack of common understanding of the same data between the sender and receiver. Explicit semantic relations among terms in digital dictionaries, such as ontologies can enable the meaning of a roadway concept name to be transparent and unambiguously understood by computer systems. However, due to the lack of an effective automated method, current practices of identifying these relations hardly rely on a manual process of knowledge acquisition from domain experts or text documents which is laborious and time-consuming. This paper presents a novel methodology that leverages recent advances in Natural Language Processing (NLP) techniques to extract English-American roadway terms used in different government agencies and their semantic relations from roadway design manuals and specifications. The proposed method includes the following three stages: (1) implementing NLP techniques to detect commonly used technical terms from the highway corpus; (2) utilizing machine learning to learn the semantic similarity among roadway terms using their context data in the corpus; and (3) developing a classification algorithm to identify semantic relation types among technical terms. The key merit in this technique is the automated identification of semantic relations among heterogeneous roadway terms from design guidebooks without reliance on other existing hand-coded semantic resources. The proposed methodology was evaluated by conducting an experiment comparing the automatically-identified synonyms by the proposed system with a human-constructed golden standard dataset obtained from Wikipedia. The result shows that the proposed model achieves a precision of over 80 percent.