Automatic generation of civil engineering lexicon to support natural language based data retrieval

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

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

ABSTRACT

Digital datasets are presented only in computer-readable formats and they are mostly complicated. In order to accurately extract a required subset of data, end users need to have deep understanding of the structure of the data schema, the meaning of each data entity and a query language. Thus, to truly facilitate the reuse of digital project data, a computational platform is needed to allow users to present their data needs in natural language. One of the critical requirements for a computer to perform this task is the ability to understand and interpret users' natural language inputs where keywords are a basic linguistic component. This research aims to collect technical terms commonly used in the civil infrastructure domain and develop a semantic similarity model that can measure the meaning relatedness/similarity between terms. Natural Language Processing (NLP) techniques and C-value method are used to automatically extract terms from text documents. A machine learning model called Skip-gram model is then employed to learn the semantic relatedness between technical terms using the unlabeled highway corpora as the input data. The input corpus includes over 15 million words mainly collected from roadway design guidelines across the U.S. The model is evaluated by comparing the mapping results performed by a computer and a human.

[1] Ph.D. Student, Department of Civil, Construction and Environmental Engineering, Iowa State University. Ames, IA 50011, United States. E-mail: ttle@iastate.edu.

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

Keywords: Civil infrastructure project, Landxml, Data retrieval, Natural language interface, NLP, Vector space model

INTRODUCTION

Neutral data standards have been widely accepted as the solution to the interoperability issue in the construction industry. Several open standards have been proposed, ranging from solely relying on syntactics using Express Modeling Language such as Industry Foundation Classes (IFC) (buildingSMART 2015) or LandXML (landxml.org 2015) to semantics-rich ontologies such as e-COGNOS (Lima et al. 2005). These standardized data models consist of rich sets of data elements covering various business processes and disciplines. However, since a specific data exchange scenario needs only a subset of data, hence neutral data standards alone are insufficient to facilitate seamless digital data exchange among project stakeholders (Froese 2003; East et al. 2012). As querying data on those data schema which are large and complicated the end user is required to have considerable programming skills and properly understand the structure and the meaning of each entity or attribute included in the source data schema. Data driven decision making based on a wrong extracted dataset would likely lead to a wrong decision. Therefore, there is a need for the formal definitions of schema subsets determining the right data for specific transactions. The availability of these model views will underpin the extraction of data from complicated sets of data generated from the AEC industry.

The MVD development is a process of matching required data items for a specific domain to the data names in the neutral data schema. A considerable amount of research efforts has been made in both the building and transportation sectors with the same ambition to define subsets of data for various business processes. One of these efforts is the Construction to Operation Building Information Exchange (Cobie) project (East 2007) which is now becomes a part of a variety of national standards and guidelines for projects using Building Information Modeling (BIM), for instance UK COBie 2.4 (Nisbet 2012), National BIM Standard-United States Version 3 (NBIMS-US) (National Institue of Building Sciences (NIBS) 2015), GSA-

BIM Guide (U.S. General Services Administration (GSA) 2011). This research identified IFC data elements that are generated in the design and construction phases required to be transferred to the asset management phase. The civil sector also is going on this trend with several model views of the Landxml schema has been being defined. The examples of these include the InfraModel project carried by the Technical Research Center of Finland aims to specify subsets of LandXML schema for several transportation projects and this specification has become the Finish national application specification (inframodel.fi 2014). Even though a considerable number of research have been made, but these are still limited to a large demand from the industry. This is because the current method for developing model view definition is based on a manual basic which is time consuming (Venugopal et al. 2012b; Eastman 2012; Hu 2014). The business processes are dynamic and tend to change over time. To adapt to the changes from industry practices, these model view are required to be tailored. Therefore, there is a need to change the current practice of model view definition from the ad-hoc approach to a more rigorous methodology (Venugopal et al. 2012b).

In this task, NLP techniques, text based ontology learning methods and the proposed vector space model in task 2 will be utilized to construct a highway machine-readable dictionary. The dictionary will organize terms in a lexical hierarchy through semantic relations such as synonym, attribute, hyponym and hypernym. Since this knowledge base is represented in a machine readable manner, it allows for the automation of variety of knowledge works such as reasoning for code compliance checking or constructability review, information retrieval, and word sense disambiguation in digital data exchange. However, like other domains, the highway industry stores knowledge in text documents which are readable to only human. This task aims to transform highway domain knowledge in natural language into a machine-readable format.

In order to eliminate the human-reliance MDV procedure, the automated method and tool are demanded. Those program needs to automatically match user data needs to the data labels in the the integrated data space. Using keywords to extract data from a database is more favorable to the end user thanks to the ease of use. This can properly done without mismatching if computer can understand the meaning of user requirements instead of string sequence based matching. Technical keywords/terms in domain specific documents implicitly refer to something that only expert in that field can correctly understand. For example, the term roadway type, in general context it can be a classification system of roadway in terms of material, function, location, etc. But in the highway context, it refers to roadway functional classification. Another example about the use of different terms for the same thing is that the 'vertical alignment' which is the long section and is the longitudinal centerline of road way have variety of terminology including 'profile', 'crest', 'grade-line'. Different DOTs may have their own vocabulary system that usually attached as glossary in their documents. different terms may be used to describe about the same thing (many different way to describe the same thing). In this case, computer is unable to to exact to map term between data sender and data receiver during the data exchange process if the algorithm is only relying on the data label/name. Due to this reason the current practices of MDV generation are still a manual process.

Recent advances in computer science with considerable improvements have enabled computer to understand human-readable format. This is thanks to the achievement in semantic measure related research which provide infrastructure for computer to present technical keywords in numeric format which can be understood by computer. A large number of methods have been proposed ranging from statistical method to machine learning such as. Distributional model is one of the most common and has been widely used. These achieve high accuracy. The availability of these offer potentials tools for the construction industry to enhance the manual work matching technical keywords in a specific domain to the open data schema.

To fill the gap above, this research aims to propose a novel model that can be used to measure the semantic similarity between technical terms in the civil infrastructure domain. In order to achieve that goal, Natural Language Processing (NLP) techniques and

C-Value method [9] are employed to process domain-specific guidelines and extract technical terms commonly used in the civil sector. A matching algorithm implementing the result from the previous step is developed to automatically look for the most nearest entities and attributes in the Landxml schema for a certain keyword. The proposed semantic similarity model and the data mapping algorithm are evaluated by comparing the automatic retrieved data with the manual results from a human for performance assessment.. The framework was complied into a Java package and full lexicon datasets which is available at https://github.com/tuyenbk/mvdgenerator.

RELATED RESEARCH

This research employed a hybrid approach which combines a series of techniques related to text analysis and semantic similarity measurement to semantically match user's input keywords to the data entities in the sources schema. Each technique is meant to support each phase of the proposed methodology. The details of research methodology will be presented in the section 3 below. This section presents a the state-of-the-art regarding partial model extraction in the construction industry and a brief introduction to the techniques deployed in the research framework.

Semantic data label matching

In current practices, data input to support a certain data analytics process usually come from multiple resources. These data are stored in different formats and are based upon different vocabularies systems. These inconsistency restricts the ability of data integration and likely leads to semantic ambiguities. In the ontology based data integration and exchange mechanism, ontology serves as a domain data schema. To allow for data exchange or integration, target and source ontology are required to be matched to each other. Matching is the process of find corresponding relationships (e.g. sameAS, isA, etc.) among semantic entities (concept, words, sentences, instances, etc.) between these ontologies (Harispe et al. 2015). These relations are found thanks to the semantic measures which determine the degree of relatedness between concepts (Harispe et al. 2015).

Lacking of an extensive machine-readable dictionary for the civil infrastructure domain

Digital dictionaries, which present definitions of terms in a machine-readable manner, are critical for computer to perform knowledge works such as interpreting users' intention or understanding the meaning behind human-oriented inputs. However, there is still a shortage of such an extensive dictionary for the civil engineering domain. WordNet (Miller 1995) (Miller 1995), which is one of the largest lexicons with over 117,000 synsets, is still generic and not suitable for the highway domain. A few construction domain specific semantic resources have been proposed, for example the Civil Engineering Thesaurus (CET) (Abuzir and Abuzir 2002) (Abuzir and Abuzir 2002), e-Cognos (Wetherill et al. 2002) (Wetherill et al. 2002), and buildingSMART data dictionary (ISO 12006-3) (buildingSMART 2016)(buildingSMART 2016) . Of these knowledge bases, the buildingSMART dictionary is a pioneer semantic database with a long development history of over two decades by the international collaboration of buildingSMART Norway, Construction Specifications Canada (CSC), U.S. Construction Specification Institute (CSI), and STABU Foundation (Hezik 2008) (Hezik 2008). Like other construction specific digital dictionaries, IFD is mainly hand-coded and time consuming; the vocabulary set covers limited number of concepts. Therefore, there is a demand for a computational technique that can automatically develop and maintain these digital dictionaries to keep up with the increasingly arising of new terms.

Lacking of effective semantic mapping algorithms for handling the data ambiguity issue

In the construction industry, research efforts are currently focusing on standardizing the data structure format, there are few research have been done to deal with the issue of sense ambiguity. Zhang and El-Gohary (2015) (Zhang and El-Gohary 2015) proposed an algorithm called ZESeM aiming to match a certain keyword to the most semantic nearest IFC entity. The algorithm includes two sequential steps including term-based matching and semantic relation based matching. Since the algorithm accepts matches from the label-based matching step, disambiguation still remains in cases in which the same word form is used for different senses. In addition, ZESeM relies on Wordnet which is a generic lexicon, the applicability would be limited. Lin et al. (2015) (Lin et al. 2015) developed a IFD based framework for BIM information retrieval. IFD Library (International Framework for Dictionaries library), which is developed and maintained by the international buildingSMART, is a dictionary of BIM data terminology that assigns synonyms the same ID. The integration or exchange of data using IDs rather than data names would eliminate semantic mismatch. However, since IFD is a hand-made electronic vocabulary, constructing this e-dictionary is time consuming and therefore it is still very limited to large collection of terms in the construction industry.

Natural Language Processing

NLP is a collection of techniques that can analyze and extract information from natural language like text and speech. The major applications of NLP include translation, information extraction, opinion mining (Cambria and White 2014). These applications are supported by a combination of several techniques such as Named Entity Recognition (NER), Part-of-Speech (POS) tagging (Toutanova et al. 2003; Cunningham et al. 2002), tokenization (or word segmentation) (Webster and Kit 1992; Zhao and Kit 2011), relation extraction, sentence parsing, word sense disambiguation (Lesk 1986; Yarowsky 1995; Navigli 2009), etc. NLP methods can be classified into two main groups: (1) rule-based and (2) machine-learning (ML) based methods. Since the early group, rule-based NLP, was based solely on hand-coded rules, these systems are not able to cover the complicated set of human grammatical system (Marcus 1995) and, therefore, do not perform well. The current trend in NLP research is the shift from rule based analysis to statistical ML based methods (Cambria and White 2014). ML models are able to learn patterns from training examples to predict the output, hence they are independent to languages, linguistic grammars and consequently reduce human resources cost (Costa-Jussa et al. 2012).

Methods for automated measuring semantic similarity

Semantic measurement, which aims to evaluate the similarity or relatedness between semantic units (words, phrases, sentences, concepts, etc.) (Harispe et al. 2015) (Harispe et al. 2015), is one of the main NLP related research topics. The two major approaches for semantic measure include (1) dictionary-based method and (2) distributional method (Harispe et al. 2013) (Harispe et al. 2013). The former method relies on a digital dictionary that consists of terms organized in a lexical hierarchy of semantic relations such as synonym, attribute, hypernym/hyponym, etc. Computational platforms (e.g., information retrieval) built upon such dictionaries are able to fast measure the semantic similarity by computing the distances between words in the hierarchy. Hence, this method would be an ideal solution when digital dictionaries are available. However, digital dictionaries are typically hand-crafted; they are therefore not available to many domains (Kolb 2008) (Kolb 2008). The latter major method for estimating word similarity is based on the distributional model which represents meanings of words through their contexts (surrounding words) in the corpus (Erk 2012) (Erk 2012). A distributional model stands on the distributional hypothesis that states that two similar terms would occur in the same context (Harris 1954) (Harris 1954). The outcome of this approach is a Vector Space Model (VSM), as illustrated in Figure 1, in which each vector depends on the co-occurrence frequencies between the target word with other words in the vocabulary. The similarity between semantic units in this model is represented by the distance between corresponding points (Erk 2012) (Erk 2012). VSM outperforms the dictionary-based method in terms of time saving as the semantic model can be automatically obtained from text corpus and collecting of these corpus is much easier than manually constructing a digital dictionary (Turney and Pantel 2010) (Turney and Pantel 2010).

The VSM approach has been used in the recent NLP related studies in the construction industry. For example, Yalcinkaya and Singh (2015) (Yalcinkaya and Singh 2015) utilized VSM to extract principle research topics related to BIM from a corpus of nearly 1000 paper abstracts. In addition, this approach was used for information retrieval to search for text documents (Lv and El-Gohary 2015) (Lv and El-Gohary 2015) or CAD documents (Hsu 2013) (Hsu 2013). The increasingly number of successful use cases in the construction industry have evidently demonstrated the promising of the VSM in identifying the semantic

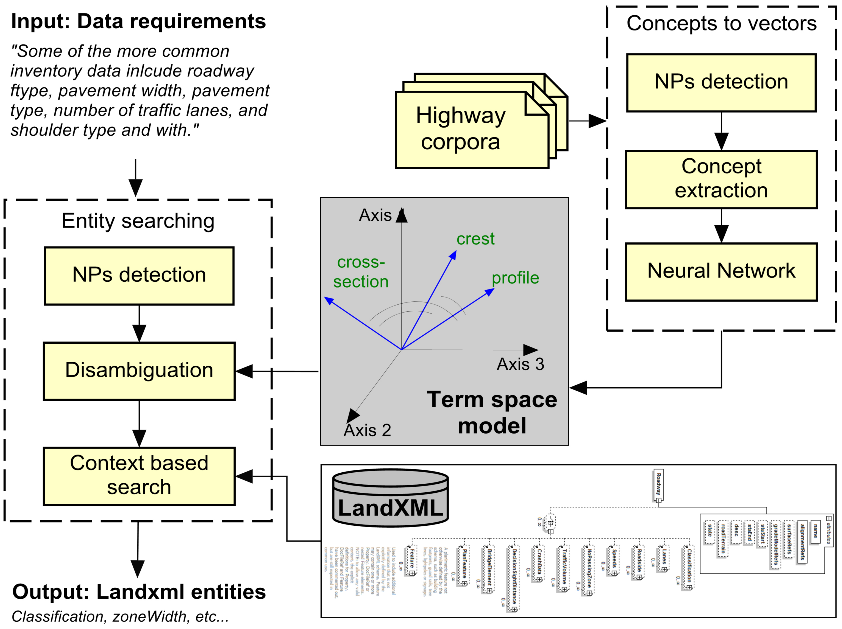

FIG. 1: Overall architecture for automated generation of MDV

similarity between technical terms in order to develop an advanced tools for handling data stored in natural language documents generated through the project life cycle.

Among the methods to develop VSM, Skip-Gram model (Mikolov et al. 2013), which is an un-supervised machine-learning model, outperforms other statistical computational methods in various performance aspects such as accuracy and degree of computational complexity (Mikolov et al. 2013). This machine-learning model learns the semantic similarity between two technical terms through their context similarity. The outcome of the training process is a set of representation vectors for technical terms.

OVERALL ARCHITECTURE

Figure 1 presents the proposed method for automated generation of model views. The framework is consisted of two components that are: (1) a highway term space model, and (2) a semantic searching algorithm. The first module previously presented by the authors in, aims to extract highway related technical concepts from the highway corpora and transform concepts into vectors representing their meanings by employing a neural network (NN) model and a set of NLP techniques. In this paper, the model is extended with larger training datasets and post-processing to reorganize terms in categories which improves the semantic data searching algorithm. Using this concept vector space, synonyms or associated concepts can be determined based on the distance or angle between vectors. The purpose of the second module is to semantically search for LandXML entities and attributes based on the natural language data requirement inputs. In order to achieve this objective, NLP techniques firstly are applied on the natural language data requirements to extract keywords that representing what types of data needed to be transferred to the data receiver. A proposed algorithm then is utilized to disambiguate the meaning of extracted required data keywords based on the term space model and search for equivalent entities or attributes included in the LandXML data schema. The following sections respectively presents the process of building the highway term space model and the searching algorithm along with details on which methods/tools utilized.

HIGHWAY TERM SPACE MODEL

This section presents the extension of the H-VSM developed by the authors' previous work with the extending of training datasets and post-processing of the vector space model. The ultimate goal of this module is to build a model that can support the disambiguation task. For disambiguation, there are several methods including thesaurus based, ontology based and distributional method. The first two methods required a full lexicon or ontology including concepts description for all aspects/disciplines in the highway industry. These methods would be ideal for the disambiguation task if domain related thesauruses are available. However, since building up those dictionaries required a huge amount of empirical work, they are still limited. Wordnet (Miller 1995) which is one of the largest lexicons available containing 117,000 synsets, but it is generic and is not suitable for the highway domain. For this reason, this research employs the distribution method which is based on an unsupervised machine learning method to train unlabeled data and learn the meaning of words by analyzing the context of words. For this reason, this research employed an unsupervised machine learning method called Skip-gram, proposed by (Mikolov et al. 2013), to train unlabeled data and learn the meanings of words by analysing their context words. The process of developing the H-VSM includes the following steps: (1) text document collection, (2) technical term extraction, and (3) semantic similarity training. The sub-sections below discuss the detailed procedures for each step.

Data collection

As mentioned earlier, the H-VSM was built using a machine learning model which use a text corpus as the training data. To support the learning of highway related technical term a highway corpus was collected in this study was built upon a number of documents from multiple sources including textbooks, and highway engineering manuals from federal Department of Transportation and from 22 state DOTs. The focus of highway corpora in this this research were on three project phases including design, construction and asset management. Technical terms in a guidance document in the engineering field are organized in various formats such as plain text, tables, and equations. Since tables and equations are not yet supported by the state-of-the-art NLP techniques, they are removed from the text corpora. The result of data collection is a plain text corpora consisting of 16 million words. This dataset is utilized to extract highway related technical terms which are then trained and converted into vectors.

Pre-processing and multi-word term candidate extraction

The collected corpus above was processed to extract term candidates. Linguists argue that a technical term is either a noun (e.g., road) or a noun phrase (NP) (e.g., right of way) that frequently occurs in the domain text documents (Justeson and Katz 1995) (Justeson and Katz 1995). Therefore, nouns and NPs are good term candidates. The linguistic process, as illustrated in Figure Figure 2. To support the generation of term candidates, a number of NLP techniques including sentence segmentation, tokenization and Part of Speech (POS) tagging will be applied on the corpus to assign tags (adjective, noun, pronoun or verb) to each of the words. The OpenNLP library is used to perform this task. The following rules, which are modified from the filters to extract multi-word terms suggested by Justeson and

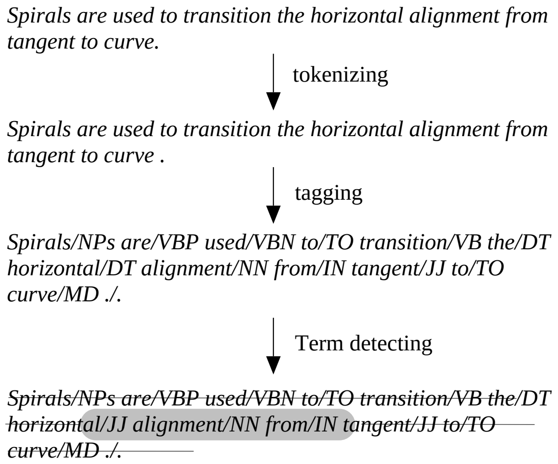
Spirals are used to transition the horizontal alignment from tangent to curve.
tokenizing
Spirals are used to transition the horizontal alignment from tangent to curve .
tagging
Spirals/NPs are/VBP used/VBN to/TO transition/VB the/DT horizontal/DT alignment/NN from/IN tangent/JJ to/TO curve/MD ./.
Term detecting
~~Spirals/NPs are/VBP used/VBN to/TO transition/VB the/DT horizontal/JJ~~ alignment/NN from/IN tangent/JJ ~~to/TO curve/MD ./.~~

FIG. 2: Linguistic processing procedure to detect technical terms

TABLE 1: Term candidate filters

Pattern	**Examples**
(Adj|N)*N	road, roadway shoulder, vertical alignment
(Adj|N)*N Prep (of/in) (Adj|N)*N	right of way, type of roadway
(Adj|N)* 'and/or' (Adj|N)*N	vertical and horizontal alignment

Note: |, * respectively denotes 'and/or', and 'zero or more'.

Katz (1995)(Justeson and Katz 1995), will be applied to extract single-word and multi-word term candidates.

The first two filters directly detect noun phrases that occur separately, and the third filter is to count for cases where multiple terms are represented in conjunctions (e.g., ='vertical and horizontal alignment'). For each instance of conjunction, an extra processing will be applied to break it into individual terms. For example, the conjunction 'vertical and horizontal alignment' will become 'vertical alignment' and 'horizontal alignment'. This division process determines the main part ('alignment') which is shared by two terms and the dependent parts ('vertical' and 'horizontal'). This research will use the Stanford Dependencies Parsing tool , which is able to analyze dependencies between sentiment units, to split conjunctions phrases into separate phrases.

In order to avoid the distinguishing between syntactic variants of the same term, for example ’roadway’ and ’roadways’, term variants will be normalized. The following are three types of syntactic variants and the proposed normalization methods.

- Type 1 - Plural forms, for example ’roadways’ and ’roadway’. The Porter stemming algorithm (Porter 1980) (Porter 1980), which can assist the automated removal of suffixes will be applied on the corpus before extracting NPs.
- Type 2 - Preposition noun phrases, for example ’roadway type’ and ’type of roadway’. In order to normalize this type of variant, the form with preposition needs to be converted into the non-preposition form by removing the preposition and reverse the order of the remaining portions. For example, ’type of roadway’ will become ’roadway type’.
- Type 3 – Abbreviations, such as AADT. A linguistic rule-based method suggested by Nenadic et al. (2002) (Nenadić et al. 2002) will be used to determine the full term for each abbreviation. This method suggested the following abbreviation definition patterns: (1) left definition pattern – NP (Abbreviation), for example Annual Average Daily Traffic (AADT); and (2) right definition pattern - (Abbreviation) NP, for example (AADT) Annual Average Daily Traffic.

The candidates with the frequency below the threshold of 50 were discarded from the list.

Multi-word term candidate raking and selection

Multi-word term definition varies between authors. in this research multiple term refers to noun phrases consiting of multiple words inwhich their meaning is not simply cobining meaning from their individual words. and there a lack of formal rules for defining multi-word term (Frantzi et al. 2000). The termhood, which represents the degree that a linguistic unit is a domain-technical concept, of the extracted candidates will then be computed based on their frequencies in the corpus; and the top candidates will be selected. There are a number of methods for termhood evaluation such as TF-IDF (Sparck 1972; Salton and Buckley 1988)

(Sparck Jones 1972; Salton and Buckley 1988), C-Value (Frantzi et al. 2000) (Frantzi et al. 2000), Termex (Sclano and Velardi 2007) (Sclano and Velardi 2007). Among these methods, Termex outperformed other methods on the Wikipedia corpus, and C-Value was the best on the GENIA medical corpus (Zhang 2008) (Zhang et al. 2008). This result indicates that C-value method is more suitable for term extraction from a domain corpus rather than a generic corpus. For this reason, the C-value has been widely used to extract domain terms in the biomedical field (Ananiadou et al. (2000) (Ananiadou et al. 2000), Lossio-Ventura et al. (2013) (Lossio-Ventura et al. 2013), and Nenadic et al. (2002) (Nenadić et al. 2002)). Since the corpus used in this research will be mainly collected from technical domain documents, thus C-value would be the most suitable for termhood determination. The C-value measure, as formulated in Equation 1, suggests that the longer a noun phrase 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.

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

where:

a is a candidate noun phrase

f is the frequency of a in the corpus

Ta is the set of extracted noun phrases that contains a

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

To calculate the recall value, expert is required to final all true terms from the corpus (Frantzi et al. 2000). with the large corpus in this research, it is impossible to do this task.

The ranked list of term candidates will be further refined; and only ones that have C-value greater than a threshold value will be accepted as technical terms. The selection of

TABLE 2: Examples of extracted terms and evaluation

Term	**Termhood**	**real term?**
sight distance	9435.314	yes
design speed	9052.556	yes
additional information	1829.0	no
typical section	1801.0	yes
basis of payment	1762.478	no

this boundary value will be evaluated using the precision and recall measures (see Equation 2 and 3). Precision shows the percentage of correctly extracted terms in the extracted list; and recall measures the percentage of technical terms in the corpus that are extracted. Experts in the highway industry will be invited in the evaluation phase to determine if a technical term is correctly recognized; and based on this, precision and recall values will be determined. The priority in this research is to minimize the number of missing domain terms (high recall), thus the threshold value will tend to be low. However, if the term list gets longer, there is a higher chance that non-technical terms are selected (low precision). Thus, domain experts' opinion will be used to determine a threshold value that balances between these two criteria. The list of extracted ranked candidate with the c-value below the value of 0 was discarded from the list which was then manually evaluated true terms and non-terms. term would provide information for a certain concepts. for example, project location, this is a compound phrases of project and location, would provide one attribute of project is location. result: 8922 terms with c-values greater 0.

Preparing training dataset

The tagged and stemmed text corpora resulted from Task 1 will be reused in this task to serve as a data source for developing the semantic similarity model. The Skip-Gram model requires a set of training data in which the input data is a linguistic unit (word or term) and the output data is a set of context words. In order to collect this training dataset, the unannotated text corpora will be scanned to collect instances of terms and

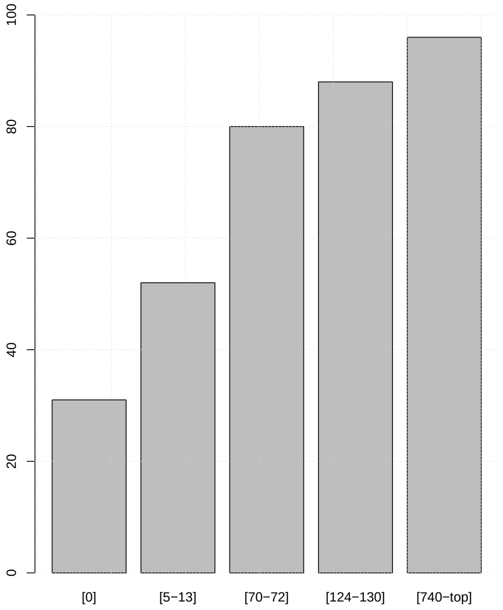

FIG. 3: Precision of term extraction

their corresponding context words. Each occurrence of a technical term will correspondingly generate a data point in the training dataset.

Before collecting the training dataset, an additional step will be performed 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. If multi-word terms are just simply replaced with unique blocks of words, the scanner will not be able to collect nested terms involved in longer terms. To allow the scanner to collect both full and nested terms, a new single unit that is compiled by connecting all individual words of a term will be added right after the last portion of the term. In the sample sentence below, the new unique unit 'signal_controlled_intersection' is added after the single-word term 'intersection'.

Before: *"A signal-controlled/JJ intersection/NN depends/VB on/IN traffic/NN signals/NN..."*

After: *"A signal-controlled/JJ intersection/NN signal_controlled_intersection/NN*

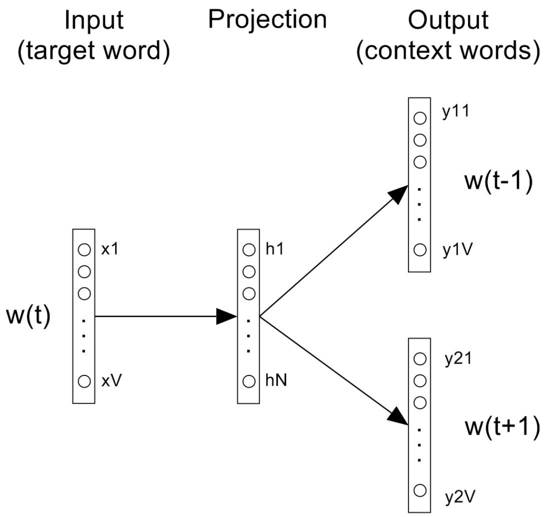

FIG. 4: Skip-gram model

depends/VB on/IN traffic signals/NN... "

The adjusted highway corpus is scanned through to find the context words for every occurrence of the technical terms in the vocabulary. The number of context words to be collected is dependent on the window size that limits how many words to the left and to the right of the target word. In the example sentence below, the context of the term 'roadway' with the context window size 10 will be the following word set bike, lane, width, on, a, width, no, curb, gutter. Any word in the stop list (a list of frequent words in English such as 'a', 'an', 'the' that have little meaning) will be neglected in the context set. If the target word is a multi-word term, the set of context words will not include its member words.

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

Data training

The semantic similarity will be trained using the word2vec of the Apache Spark MLlib package (Apache.org 2016), an emerging open-source engine, developed based on the Skip-gram neural network model (Mikolov et al. 2013) (Mikolov et al. 2013). This tool includes a module that allows for automatic collection of training data from a text corpus. However, the algorithm only supports the collection of data points from single words, this module

TABLE 3: Skip-gram model parameters

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

will be modified using the proposed method described in Task 2a so that multi-word terms and their nested terms can be obtained. The intended parameters used for the training model are presented in Table 3. These values are mainly based on suggestions from the literature. To eliminate data points with low frequency of occurrence that are unlikely to be technical terms, word2vec includes the parameter of minimum occurrence frequency. Any vocabulary with the rate lower than the limit will be ignored. Radim Rehurek, a machine learning consultant company, suggests a range of (0-100)* depending on the data set size. Setting this parameter high will enhance the accuracy, but many technical terms will be out of vocabulary. The preliminary study based on the preliminary corpus with only several millions of words shows that with the frequency of 20, there are very few non-technical terms involved in the training dataset. Hence, with the larger dataset to be collected, this parameter can be higher and up to around 50. The second important parameter is layer size which determines the number of nodes in the hidden layer. This parameter highly affects the training accuracy and processing time. A larger layer size is better in terms of accuracy but this will be paid off by the running time. Since this research aims to develop a model that can be used for other information retrieval research, the accuracy is the first priority. This parameter may range from 10 to hundreds; in this research, it is expected to in be the range of 100-500. The final major parameter is the context window size. Google suggests the size of 10 for the Skip-gram model. In future work, these parameters would be subject to be change so that the best model can be achieved. Figure 3 presents the term space model developed from the training process with the parameters are 50, 300 and 10 respectively. In this model, each technical term collected from technical documents is represented as a

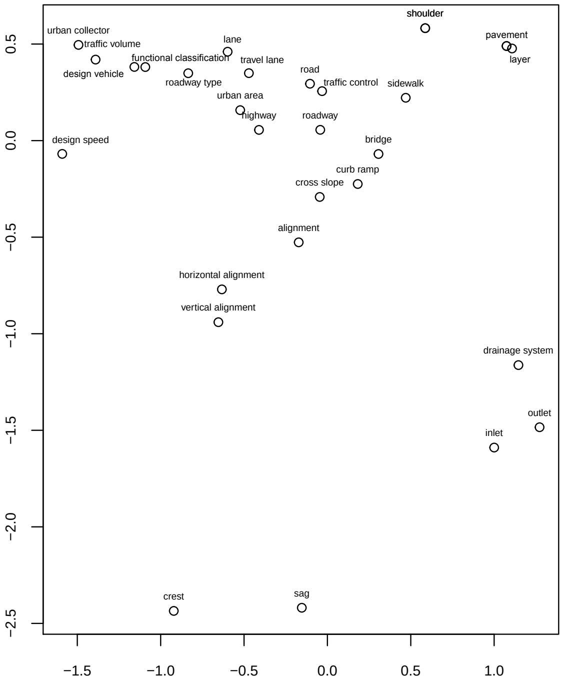

FIG. 5: Highway term space model (H-VSM)

vector in a high dimensional space; and the distance between them represents the semantic similarity. The preliminary term space presented in this paper consists of more than six thousand technical keywords. Since the vector space is a multi-dimensional space according to the size of hidden layer. In order to illustrate, present space in 2D graph, PCA (Principle component Analysis) was used to reduce the size to 2 dimensions.

The similarity between terms can be measured by the angle between two work representation vectors or the distance between two word points. The following shows two measures of word sense similarity.

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

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

Where: n is the hidden layer size.

HIGHWAY LEXICON CONSTRUCTION

A knowledge base typically includes terms and relations. The specific objective of this

TABLE 4: Examples of top nearest terms

Term	Nearests	Cosine	Rank
roadway	highway	0.588	1
	traveled-way	0.583	2
	roadway-section	0.577	3
	road	0.533	4
	traffic-lane	0.524	5
	separating	0.522	6
	adjacent-roadway	0.519	7
	travel-way	0.517	8
	entire-roadway	0.513	9
	...	...	...
	roadway-shoulder	0.505	12
	roadway-cross-section	0.491	18
	undivided	0.452	37
	mainline-roadway	0.450	42

task is to detect relations and based on that rearrange the vocabulary constructed in Task 1. The core relations of an ontology can be classified into the following types: 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). A method integrating linguistic rules, machine learning and the term similarity model developed in Task 2 will be proposed to extract those relations. Among these relations, hypernym, hyponym and attribute can be detected independently using linguistic rules and/or statistical methods. In contrast, since synonym recognition is well known as the process of evaluating the sharing of common attributes, hypernyms, and hyponyms, this process will rely on the results from the detection of other relations. This task will first detect the following relations (hypernyms, hyponyms and attributes) and then use them as features to find synonyms.

the list of semantic nearest words generated from the semantic similarity model contain a list of various type of relations. this research consider the following three groups of relations including (1) attribute, (2) hyponym and (3) synonym/sibling and functional relation. For example, in the example presented in Table 4, true synonyms are highway (1), traveled-way (2) or road(4); roadway-section (3), roadway-shoulder (12) are attributes of the roadway concept; and adjacent-roadway (7), undivided (37) are hyponyms which showing differnt types of roadway.Tthe following sections present the procedure followed to reorganized the extracted list of nearest words into the three categories.

Algorithm 1 Near term 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
```

Attribute and hyponym patterns

The list of near words consists of various types of relationships such as attribute (lane-width), hypernyms-hyponyms (roadway-facility), synonyms (sag-profile) and other functional related terms (lane-bike). The list of nearest words will be reorganize into three categories: (1) attributes, (2) hyponyms and (3) other near term. This task aims to determine the typical attributes for the highway project related concepts using a hybrid method.

TABLE 5: Patters to extract attributes

Relation Pattern	Example	
Attribute	Noun of Target	the width of the road
	Target Noun	road width, project cost
Hypernym-hyponym	Noun Target	vertical alignment isA alignment

Attributes, also known as the ‘property-of’ relation in ontology learning, is critical to describe a real-world concept. This research will follow the method suggested by Lee et al. (2013) (Lee et al. 2013) to extract attributional information from the highway corpus collected in Task 1. The corpus will be scanned to collect word sequences that satisfy a certain feature pattern. This research will perform an extensive review on pattern-based attribute extraction research to collect the list of feature patterns. Examples of these patterns are presented in Table 3. For example, with pattern 1, we can infer that Noun1 is a good attribute candidate of concept Noun2; and Noun2 is an attribute of Noun1 in the pattern 2. For the second group

•

After the candidate list is refined, the frequency of occurrence for each candidate will then be used to compute the degree that term ’a’ is an attribute of concept ’c’. If ’a’ is a typical attribute of ’c’, it should frequently occur in the corpus. Each concept ’c’ will correspondingly have a list of attribute candidates (called list A) and their frequency of occurrences. The likelihood that ’a’ is an attribute of concept ’c’ is estimated using the normalized probability formula (see Equation 2). The attribute candidates for each concept will be ranked by the likelihood measure and the top list over a threshold value will be accepted as typical attributes.

$$P(a|c) = \frac{n(c,a)}{\sum_{a*\in A} n(c,a)} \tag{2}$$

TABLE 6: Examples of top nearest terms

Term	Relation Group	Nearests	Cosine	Rank
roadway	Synonym	highway	0.588	1
		traveled-way	0.583	2
		road	0.533	4
		traffic-lane	0.524	5
		travel-way	0.517	8
	Attribute	separating	0.522	6
		roadway-section	0.577	3
		roadway-shoulder	0.505	12
		roadway-cross-section	0.491	18
	Hyponym	adjacent-roadway	0.519	7
		entire-roadway	0.513	9
		undivided	0.452	37
		mainline-roadway	0.450	42

Synonym/sibling and functional relation recognition

the previous proposed method are used to recognize the first two group. the remain words will fall into the final group. however, this list include long list of terms. and 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 divide the remaing list into multile layers based on their similarity score. the model can provide a list of nearest words and terms. this list is needed to filter to selected the most nearest ones. this research employ the k-mean clustering algorithm (MacQueen 1967) to discard unclosed terms. the nearest list is clustering into three distinguished groups based on their similarity score. the last group is removed from the nearest list.

EVALUATION

An evaluation experiment was conducted to evaluate the context-aware searching algorithm. In this experiment, a graduate student was asked to read a randomly selected document which contains data requirements for a specific data transaction. The duty of the student was to manually identify a set of Landxml entities that would fulfill the data requirements. Meanwhile, a prototype built upon the developed algorithm was applied to automatically generate the subset of data from the same document as the student used. The results from the two methods were used to calculate the accuracy of the searching algorithm.

in order to evaluate the performance of the hvsm model in most related term searching task, this research utilized the intrinsic evaluation method which is based on the comparison between computational algorithm and a gold standard. gold standard is a manually developed dataset of word synonyms or sharing common characteristics. there are several gold standards for evaluating semantic similarity tasks such as WordSim(WS)-353 (Finkelstein et al. 2001), MEN (Bruni et al. 2012) and SimLex-999 (Hill et al. 2014). Of these, SimLex is the nearest to this research evaluation purpose since it provide predefined list of synonyms while the latter datasets generic similarity including functional relations bothm similarity and relatedness. correlation between algorithm and human assigned rating However, these standards consist no technical terms in the highway section. This research propose the proprietary gold standard that can be used for evaluating highway specific term similarity task.the gold stand was developed based on the glossary of some technical terms in the highway section provided by Wikipedia. The glossary was inspected and final list of terms were selecting excluding terms that have no synonyms. the gold standard here include both synonym set and sibling, or hypernym-hyponym. this glossary provided plain text explanation for each term and their synonyms. the final sample of the gold stand consist of 73 terms (both signle and multi-word terms). the automatically indentified synonyms which is the top nerest word in the indentified synonym list from the algorithm was compared with the true synonyms in the gold standard dataset. the results are measured using the following three measures including recall, precision and f-measure. the evaluation result is presented in the table 7.

$$Recall = \frac{\text{number of correctly matched concepts}}{\text{total concepts}} \tag{3}$$

$$Precision = \frac{\text{number of correctly matched concepts}}{\text{total matched concepts}} \tag{4}$$

$$F - measure = \frac{2.Precision.Recall}{Precision + Recall} \tag{5}$$

Table 7 shows the evaluation result. As presented in the table, the system shows a 90 percent precision. However, the recall is relatively low accuracy, this is possibly due to the the training data size. Since the searching algorithm accuracy is highly rely on the a capacity of finding synonyms which is based on the vector space model. This model currently is based on the data training set consisting of only 10 million words. In order to enhance the accuracy, the data training set needs to be extended. Future research will be conducted to extend the training data set. compared to the previous publication by the authors, this algorithm has been improved with significantly higher accuracy. the post-processing is one reason for the outperformance of the proposed method.

CONCLUSIONS

TABLE 7: Evaluation result

Model	**Precision (%)**	**Recall(%)**	**F (%)**
50-300-15	80	42	55
75-300-15	-	-	-
100-300-15	-	-	-
100-100-15	-	-	-
100-200-15	-	-	-
100-300-15	-	-	-
100-300-5	-	-	-
100-300-10	-	-	-
100-300-15	-	-	-

Digital data has been widely generated through the project life cycle. However, the data collected and generated in previous stages are no reusable in the downstream phases. This issue is due to the interoperability when digital data from one partner is not readable or correctly understandable by the data receiver. This research develops an framework that semantically searches for desired data from the transferred data file. The framework is composed of two components including (1) a terms space model which represents highway related concepts extracted from the highway corpora in vectors and (2) a context based searching algorithm that can search for entities in the Landxml schema based on their similarity of attributes instead of string based similarity.

The framework has been evaluated by testing on a randomly selected set of input data. The result shows the accuracy of over 80 percent. The accuracy is low due to the size of the training data. Future research will be conducted to increase the data size.

This method is broad and can be applied to other business processes such as green building checking, environment checking, etc. The method is expected to significantly improve the existing ad-hoc method of model view definition development and in return leads the the removal of this bottle neck which is restricting the seamless data integration and exchange across phases of a highway construction project.

Digital project data is now widely available throughout the project life cycle in the civil infrastructure sector. However, the data collected and generated in early project development stages are not typically reusable in the downstream phases. This is due to the interoperability issue when digital data from the original data creator is not readable or correctly understandable by the data receiver. This research developed a framework that semantically searches for the desired data from a transferred data file. The framework is composed of two components including (1) a terms space model which represents highway related concepts extracted from the highway corpora in vectors and (2) a searching algorithm that can search for entities in the Landxml schema based on their semantic similarity instead of string based similarity. The framework was evaluated by testing on a randomly selected set of input keywords. The result shows the accuracy of over 30 percent. The accuracy is low due to the size of the training data. Future research will be conducted to increase the data size. This method is broad and can be applied to extract data supporting for various business processes such as green building checking, environment checking, etc. The method is expected to significantly improve the existing ad-hoc method of model view definition development and in return leads the removal of this bottle neck that is restricting the seamless data integration and exchange across phases of a highway construction project.

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).

Berard, O. B. and Karlshoej, J. (2012). "Information delivery manuals to integrate building product information into design." *CIB W78-W102 2011: International Conference*.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). "Distributional semantics in technicolor." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, Association for Computational Linguistics, 136–145.

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

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

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.

East, E. W. (2007). "Construction operations building information exchange (cobie)." *Report no.*, DTIC Document.

East, E. W., Nisbet, N., and Liebich, T. (2012). "Facility management handover model view." *Journal of computing in civil engineering*, 27(1), 61–67.

Eastman, C. (2012). "The future of ifc: Rationale and design of a sem ifc layer. Presentaion at the IDDS workshop.

Eastman, C., Jeong, Y., Sacks, R., and Kaner, I. (2009). "Exchange model and exchange object concepts for implementation of national bim standards." *Journal of Computing in Civil Engineering*, 24(1), 25–34.

Erk, K. (2012). "Vector space models of word meaning and phrase meaning: A survey." *Language and Linguistics Compass*, 6(10), 635–653.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2001). "Placing search in context: The concept revisited." *Proceedings of the 10th international conference on World Wide Web*, ACM, 406–414.

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.

Froese, T. (2003). "Future directions for ifc-based interoperability." *ITcon*, 8, 231–246.

Gale, W. A. and Church, K. W. (1993). "A program for aligning sentences in bilingual corpora." *Computational linguistics*, 19(1), 75–102.

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*.

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2015). "Semantic similarity from natural language and ontology analysis." *Synthesis Lectures on Human Language Technologies*, 8(1), 1–254.

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. (2014). "Simlex-999: Evaluating semantic models with (genuine) similarity estimation." *arXiv preprint arXiv:1408.3456*.

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

Hu, H. (2014). "Development of interoperable data protocol for integrated bridge project delivery." Ph.d., Ph.d. Copyright - Copyright ProQuest, UMI Dissertations Publishing 2014 Last updated - 2015-03-18 First page - n/a.

inframodel.fi (2014). "Inframodel, <http://www.inframodel.fi/en/>. Accessed: 2015-10-11.

Jeong, W., Kim, J. B., Clayton, M. J., Haberl, J. S., and Yan, W. (2014). "Translating building information modeling to building energy modeling using model view definition." *Scientific World Journal*, 2014 Cited By :1 Export Date: 16 February 2016.

Jiang, Y., Yu, N., Ming, J., Lee, S., DeGraw, J., Yen, J., Messner, J., and Wu, D. (2015). "Automatic building information model query generation." *Journal of Information Technology in Construction*.

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.

Katranuschkov, P., Weise, M., Windisch, R., Fuchs, S., and Scherer, R. J. (2010). "Bim-based generation of multi-model views." *CIB W78*.

Kolb, P. (2008). "Disco: A multilingual database of distributionally similar words." *Proceed-*

ings of KONVENS-2008, Berlin.

landxml.org (2015). “About landxml.org, <http://www.landxml.org/About.aspx>. Accessed: 2015-10-11.

Lee, T., Wang, Z., Wang, H., and Hwang, S.-w. (2013). “Attribute extraction and scoring: A probabilistic approach.” *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, IEEE, 194–205.

Lee, Y. C., Eastman, C. M., Solihin, W., and See, R. (2016). “Modularized rule-based validation of a bim model pertaining to model views.” *Automation in Construction*, 63, 1–11 Export Date: 16 February 2016.

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.

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.

Lin, J.-R., Hu, Z.-Z., Zhang, J.-P., and Yu, F.-Q. (2015). “A natural-language-based approach to intelligent data retrieval and representation for cloud bim.” *Computer-Aided Civil and Infrastructure Engineering.*

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2013). “Combining c-value and keyword extraction methods for biomedical terms extraction.” *LBM’2013: 5th International Symposium on Languages in Biology and Medicine*, http://lbm2013.biopathway.org/. Computer Science [cs]/Bioinformatics [q-bio.QM] Life Sciences [q-bio]/Quantitative Methods [q-bio.QM] Computer Science [cs]/Document and Text ProcessingConference papers.

Lv, X. and El-Gohary, N. M. (2015). “Semantic annotation for context-aware information retrieval for supporting the environmental review of transportation projects.

MacQueen, J. (1967). “Some methods for classification and analysis of multivariate observations.” *Proceedings of the fifth Berkeley symposium on mathematical statistics and proba-*

bility, Vol. 1, Oakland, CA, USA., 281–297.

Marcus, M. (1995). “New trends in natural language processing: statistical natural language processing.” *Proceedings of the National Academy of Sciences*, 92(22), 10052–10059.

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

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

National Institue of Building Sciences (NIBS) (2015). “National bim standard – united states version 3.” *Report no.*

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

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

Nisbet, N. (2012). “Cobie uk: Required information for facility operation.” *Report no.*, AEC3 UK Ltd.

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

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

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

See, R., Karlshoej, J., and Davis, D. (2012). “An integrated process for delivering ifc based data exchange.

Sparck Jones, K. (1972). “A statistical interpretation of term specificity and its application 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.

U.S. General Services Administration (GSA) (2011). "Gsa building information modeling guide series: 08 – gsa bim guide for facility management." *Report no.*

Venugopal, M., Eastman, C., and Sacks, R. (2012a). "Configurable model exchanges for the precast/pre-stressed concrete industry using semantic exchange modules (sem)." *International Conference on Computing in Civil Engineering*, 269–276.

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

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.

Windisch, R., Katranuschkov, P., and Scherer, R. J. (2012). "A generic filter framework for consistent generation of bim-based model views." *European Group for Intelligent Computing in Engineering, EG-ICE 2012 - International Workshop: Intelligent Computing in Engineering*. Export Date: 16 February 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, C., Beetz, J., and De Vries, B. (2013). "Towards model view definition on semantic level: A state of the art review." *European Group for Intelligent Computing in Engineering, EG-ICE 2013 - 20th International Workshop: Intelligent Computing in Engineering*. Export Date: 16 February 2016.

Zhang, J. and El-Gohary, N. (2015). "A semantic similarity-based method for semi-automated ifc exension.

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.