CSCE 2013 General Conference - *Congrès général 2013 de la SCGC*

Montréal, Québec
May 29 to June 1, 2013 / *29 mai au 1 juin 2013*

Automated Procedure for Extracting Safety Regulatory Information Using Natural Language Processing Techniques and Ontology

JunHyuk Kwon[1], Byungil Kim[2], SangHyun Lee[3], and Hyoungkwan Kim[4]

[1] MSE Student, Department of Civil & Environmental Engineering, University of Michigan, USA
[2] Postdoctoral Fellow, Department of Civil & Environmental Engineering, University of Michigan, USA
[3] Assistant Professor, Department of Civil & Environmental Engineering, University of Michigan, USA
[4] Associate Professor, School of Civil & Environmental Engineering, Yonsei University, Republic of Korea

Abstract: Traditionally, the identification of design-related hazards inherent in design drawings has been performed manually by safety experts. However, this manual approach may lead to incomplete, inaccurate, or incompatible results because of its repetitive, time-consuming, and error-prone process. For this reason, automating the safety design review process is expected to save time and reduce human interpretation errors. In this paper, we address this issue by formulating a procedure of ontology-based information extraction using natural language processing (NLP) techniques and apply it to safety review in the design phase. Specifically, construction safety requirements are identified from textual regulatory documents, and then, are converted to machine-readable format. The proposed approach was applied to extract hazard information from two different types of regulatory documents. Preliminary results demonstrate that this approach is effective in automating the hazard information extraction without the manual interpretation from safety experts.

1. Introduction

Construction workers perform their jobs under circumstances that place them at a high risk of injury, which often leads to fatal injuries. According to the Bureau of Labor Statistics (2012), a total of 721 fatal injuries were recorded in the construction sector in 2011ð the second highest number of fatal injuries among any industry sector that year. Moreover, the construction sectorõs fatal injuries rate (per 100,000 full-time equivalent workers) was 8.9, well above the all-worker average of 3.5. Pecuniary losses caused by work-related injuries including fatalities in the construction industry are also enormous. Waehrer et al. (2007) estimated the costs of work-related injuries in the U.S. construction industry, including direct medical costs and indirect losses in wage and productivity. As of 2002, the average cost per injury case was estimated to be $27,000. This estimate was almost double the average of all the industries: $15,000. In addition to the loss of money, work-related injuries have a negative impact on the performance of construction projects as a result of property and material damage, time spent on investigations, and the loss of skilled workers (Holt 2001).

A major cause of such a high fatal rate is due to design-related hazards (e.g., inappropriate height of a roof parapet) (Gambatese et al. 2008; Gambatese et al. 2005; Trewethy and Atkinson, 2003; Gambatese et al. 1997; Szymberski 1997). Behm (2005) and Gibb et al. (2004) found that approximately 50% of construction fatalities were because of the decisions made in the design phase. Thus, to maximize the safety benefits, designers must consider eliminating design-related hazards from their designs, a process known as Prevention through Design (PtD) (Howard 2008).

Designers are typically unaware of the impact of their decisions on worker safety due to the lack of adequate knowledge regarding construction planning, equipment, and methods. For this reason, the identification of design-related hazards inherent in design drawings has been manually performed by safety experts. However, this manual approach may lead to incomplete, inaccurate, or incompatible results because of its repetitive, time-consuming, and error-prone process (Zhou et al. 2012; Eastman et al. 2009). To address these issues, we have developed a framework for automatically identifying hazards inherent in design drawings (Kwon et al. 2013).

The automated framework developed in our previous study can extract safety regulatory information from Occupational Safety and Health Administration (OSHA) Regulations using Natural Language Processing (NLP) techniques, and subsequently can identify design-related hazards based on the regulatory information using Building Information Modeling (BIM) technology. However, our framework suffered low performance when applying it to other safety regulatory documents (e.g., Construction Workplace Design Solutions). Specifically, the previous procedure for extracting safety regulatory information (e.g., syntax parsing) may not be applicable to a document that is composed of different words in different sentence structures, even though the meaning of the sentences is the same as their OSHA Regulation counterparts. To overcome this issue, we aim to extract safety regulatory information from any textual document using ontology-based parsing.

The objective of this study is to design and test an automated procedure for extracting regulatory information from various unstructured documents regarding construction worker safety. In the remaining sections of this paper, the first section describes the information extraction procedure proposed in this study. The second section presents a case study applying the proposed approach to two different regulatory documents on construction worker safety. The final section concludes this paper with our findings and recommendations for future research.

2. Information Extraction

2.1. Traditional IE Procedure

The information extraction (IE) approach, which is a subfield of NLP, brings together other NLP techniques (e.g., syntax analysis) and domain knowledge to extract a textŝ concepts, and to form a structured representation based on predefined templates. Figure 1 shows the IE procedure that can be broken down into three stages: lexical analysis, syntactic analysis, and semantic analysis (adopted from Grishman 1997; Jurafsky and Martin 2009; Maynard et al. 2009).

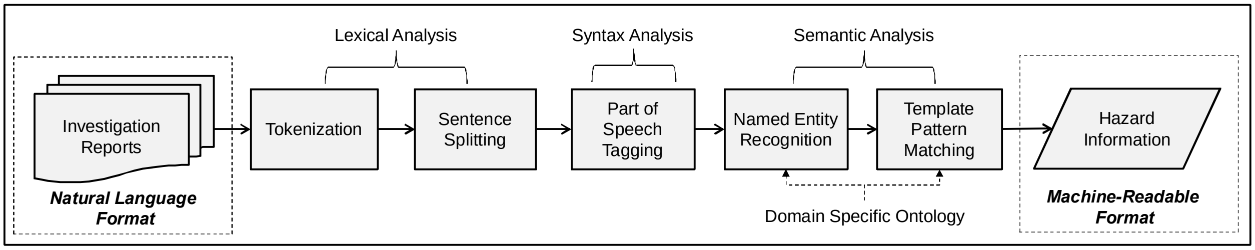


Figure 1: Procedure of information extraction

The lexical analysis is the process of splitting the text into meaningful units called tokens, such as numbers, punctuation marks (e.g., :, /, &), and words of different types (e.g., an initial capital, all upper case). The syntactic analysis is the process of analyzing a sequence of tokens (e.g., sentence) to determine its grammatical structure with respect to a grammar or principles and rules for constructing sentences in natural language. The semantic analysis is the process of assigning a given sense to the different constituents of a sentence based on a specific context in a document. The semantic (④ and ⑤) analysis is essentially related to the contextual meaning of words, whereas the lexical (① and ②) and syntactic (③) analyses are driven by a grammar, which is a set of syntactic rules that govern a language. Procedural details on the analyses are as follows (Grishman 1997; Jurafsky and Martin 2009; Maynard et al. 2009):

① Tokenization: the process of identifying individual tokens such as numbers, punctuation marks (e.g., ꝏ :, /), and words within a text.

② Sentence Splitting: the process of recognizing sentence boundaries.

③ Part-of-Speech (POS) Tagging: the process of assigning a POS (i.e., grammatical category such as noun or verb) label to each token in a sentence.

④ Named Entity Recognition: the process of identifying specific words or phrases (i.e., a series of tokens) and categorizing them (e.g., persons, organizations, locations). The gazetteer, which recognizes entities stored in its list, is often used due to its computational efficiency.

⑤ Template Pattern Matching: the process of extracting information from a sentence in regard to predefined patterns.

Throughout these processes, the information in natural language format is transformed into a tabular structure in machine-readable format that can be efficiently loaded as input variables in the step of conformance checking.

2.2. IE Procedure without a Domain-Specific Ontology

In our previous study (Kwon et al. 2013), the procedure depicted in Figure 1 was customized for extracting safety regulatory information from the OSHA Regulations, which are the most common safety regulations imposed on construction sites. The IE procedure was implemented using the General Architecture for Text Engineering (GATE), which is one of the widely used NLP toolkits.

Here, we briefly introduce the IE procedure previously developed with an emphasis on semantic analysis because the named entity recognition and template pattern matching are the most important processes in the whole procedure. The other analyses are, by their nature, conducted without the consideration of sentence structures or word choices unique in the OSHA Regulations.

- Named Entity Recognition (Process ④): A set of gazetteers was used to find occurrences of predefined names of elements in the OSHA Regulations. The gazetteers contained names of entities (e.g., building elements, work activities, and construction equipment) based on word choices in the OSHA Regulations. It should be noted that the gazetteers were manually compiled in our previous study.
- Template Pattern Matching (Process ⑤): A variety of single patterns was also manually compiled based on sentence structures of the OSHA Regulations. For this process, built-in software in GATEð the Java Annotation Pattern Engine (JAPE) transducerð was used.

On all of the tests, we could successfully extract safety regulatory information from the OSHA Regulations with 100% precision (i.e., the number of correct instances divided by the number of all extracted instances) and 94.4% recall (i.e., the number of correct instances divided by the number of instances that should have been extracted) (Kwon et al. 2013). However, compiling all of the names of elements in a text was inefficient and a template of pattern matching that fits one scenario was not directly applicable to a different scenario of template. Moreover, building all possible scenarios of patterns in sentences was difficult because there are great variations in sentence structures. For example, there is a pattern that states a sentence should be composed of whoï howï why, specifically in this order. If a sentence has a different structure (e.g., whyï howï who), it is very difficult to accurately match each word (or phase) with a concept. To overcome this difficulty in matching of lists and single patterns, this study employs a domain-specific ontology because it can provide a set of concepts within a domain, as well as a description of the relations between the concepts (Gruber 1995) and this information (i.e., concepts and their relations) can be effectively used in improving the performance of the information extraction.

2.3. IE Procedure with a Domain-Specific Ontology

In the present study, we developed an IE procedure with a domain ontology to tackle the matching problems. Here, the ontology was developed by modifying PROTON Ontology (PROTo ONtology), which is a basic upper-level ontology. To conduct IE in conjunction with the domain-specific ontology, the OntoRoot Gazetteer and the Ontology-aware JAPE transducer were used in GATE for the named entity recognition and the template pattern matching, respectively. Technical details are as follows:

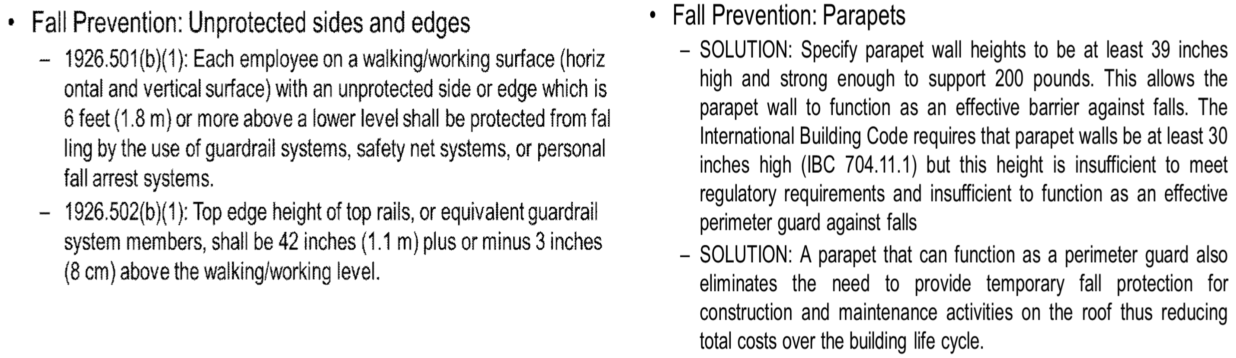
- Named Entity Recognition (Process ④): The OntoRoot Gazetteer was compiled to produce annotations linked to specific concepts (e.g., building elements, physical quantity, and construction resources) and relations from the ontology by looking up items from the ontology and matching them with the text, based on root forms. In this way, gazetteer lists are automatically created directly from the ontology resources and are then used by the subsequent processing components (e.g., JAPE transducer) to annotate mentions of classes, instances, and properties in the content.
- Template Pattern Matching (Process ⑤): The patterns are implemented in GATE as JAPE rules combining ontologies. Applying domain ontology in conjunction with the JAPE transducers can significantly simplify the set of grammars that needs to be written. On the left-hand side (LHS) of the rule is the pattern to be annotated. This consists of a number of pre-existing annotations that have been created as a result of pre-processing components (such as POS tagging or gazetteer lookup) and earlier JAPE rules. The right-hand side (RHS) of the rule gets named entities from the annotations (using labels on the LHS of the rule), then adds new annotations identifying the ontology class to the entities in the document itself.

One of the most important potentials of the above processes are their ability to automatically generate semantic annotations, which alleviate the laborious process of manually building and maintaining named entity lists (gazetteer lists) and pattern matching templates (JAPE rules) as we did in our previous work.

3. Case Study

3.1. IE Procedure without a Domain-Specific Ontology

We, first, tested the IE procedure without a domain-specific ontology, which was developed for OSHA Regulations (Figure 2a) in our previous study (Kwon et al. 2013), by applying it on Construction Workplace Design Solutions (Figure 2b), which is almost identical to OSHA Regulations in terms of semantic sameness and similarity, but it is composed of different words in different sentence structures.



- Fall Prevention: Unprotected sides and edges
  - 1926.501(b)(1): Each employee on a walking/working surface (horizontal and vertical surface) with an unprotected side or edge which is 6 feet (1.8 m) or more above a lower level shall be protected from falling by the use of guardrail systems, safety net systems, or personal fall arrest systems.
  - 1926.502(b)(1): Top edge height of top rails, or equivalent guardrail system members, shall be 42 inches (1.1 m) plus or minus 3 inches (8 cm) above the walking/working level.

(a)

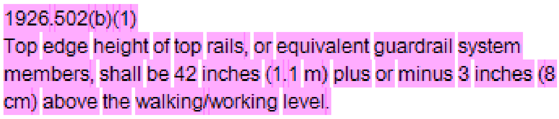
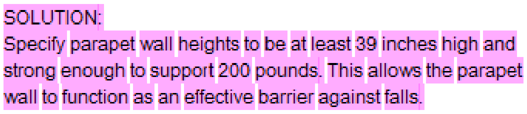
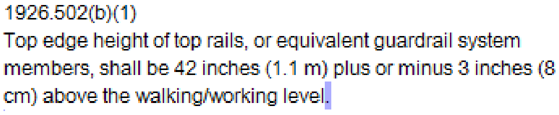
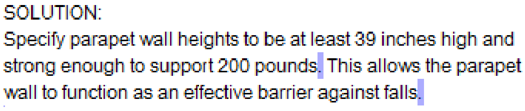
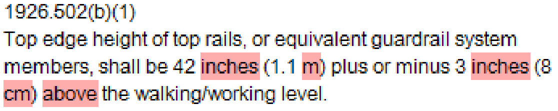
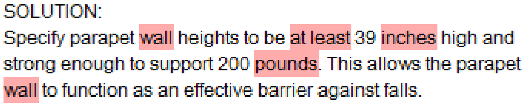
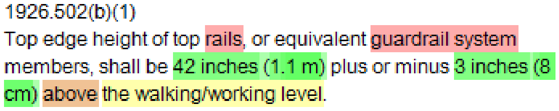
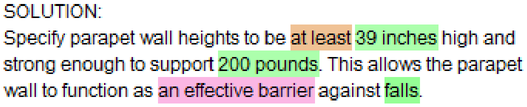
- Fall Prevention: Parapets
  - SOLUTION: Specify parapet wall heights to be at least 39 inches high and strong enough to support 200 pounds. This allows the parapet wall to function as an effective barrier against falls. The International Building Code requires that parapet walls be at least 30 inches high (IBC 704.11.1) but this height is insufficient to meet regulatory requirements and insufficient to function as an effective perimeter guard against falls
  - SOLUTION: A parapet that can function as a perimeter guard also eliminates the need to provide temporary fall protection for construction and maintenance activities on the roof thus reducing total costs over the building life cycle.

(b)

Figure 2: Regulatory documents for construction worker safety:
(a) OSHA Regulations; (b) Construction Workplace Design Solutions

Table 1 shows an example of regulatory information extraction results, process by process, using the IE procedure previously developed in Kwon et al. (2013). As can be seen from the table below, processes ①, ②, and ③ were accurately completed, whereas processes ④ and ⑤ were not. In other words, gazetteer lists (for process ④) and single patterns (for process ⑤) developed on the basis of OSHA Regulations were not applicable to Construction Workplace Design Solutions because they were composed of different words in different sentence structures. For example, although two noun phrases (e.g., ǿguardrail systemô in OSHA Regulation and ǿparapet wallô in Construction Workplace Design Solutions) had an identical concept (e.g., building element), the pattern matching rule successfully extracted ǿguardrail systemô as subjects from OSHA regulations, while ǿparapet wallô was not extracted at all from the design solutions document because the first noun phrase located ahead of the verb is defined as the subject in this rule (See the 7th row of Table 1).

Table 1: Example of information extraction from the two different documents

	OSHA Regulations	**Construction Workplace Design Solutions**
Original Sentences	1926.502(b)(1): Top edge height of top rails, or equivalent guardrail system members, shall be 42 inches (1.1 m) plus or minus 3 inches (8 cm) above the walking/working level.	SOLUTION: Specify parapet wall heights to be at least 39 inches high and strong enough to support 200 pounds. This allows the parapet wall to function as an effective barrier against falls.
①	 1926.502(b)(1) Top edge height of top rails, or equivalent guardrail system members, shall be 42 inches (1.1 m) plus or minus 3 inches (8 cm) above the walking/working level.	 SOLUTION: Specify parapet wall heights to be at least 39 inches high and strong enough to support 200 pounds. This allows the parapet wall to function as an effective barrier against falls.
②	 1926.502(b)(1) Top edge height of top rails, or equivalent guardrail system members, shall be 42 inches (1.1 m) plus or minus 3 inches (8 cm) above the walking/working level.	 SOLUTION: Specify parapet wall heights to be at least 39 inches high and strong enough to support 200 pounds. This allows the parapet wall to function as an effective barrier against falls.
③	1926.502(b)(1) Top edge height of top rails, or equivalent guardrail system JJ NN NN IN JJ NNS CC JJ NN NN members, shall be 42 inches (1.1 m) plus or minus 3 inches NNS MD VB CD NNS CD NN CC CC CC CD NNS (8 cm) above the walking/working level. CD NN IN DT VBG VBG NN	SOLUTION: Specify parapet wall heights to be at least 39 inches high and VB NN NN NNS TO VB IN JJS CD NNS JJ CC strong enough to support 200 pounds. This allows the parapet JJ RB TO VB CD NNS DT VBZ DT NN wall to function as an effective barrier against falls. NN TO VB IN DT JJ NN IN NNS
④	 1926.502(b)(1) Top edge height of top rails, or equivalent guardrail system members, shall be 42 inches (1.1 m) plus or minus 3 inches (8 cm) above the walking/working level.	 SOLUTION: Specify parapet wall heights to be at least 39 inches high and strong enough to support 200 pounds. This allows the parapet wall to function as an effective barrier against falls.
⑤	 1926.502(b)(1) Top edge height of top rails, or equivalent guardrail system members, shall be 42 inches (1.1 m) plus or minus 3 inches (8 cm) above the walking/working level.	 SOLUTION: Specify parapet wall heights to be at least 39 inches high and strong enough to support 200 pounds. This allows the parapet wall to function as an effective barrier against falls.

3.2. IE Procedure with a Domain-Specific Ontology

In order to address the mismatching issue described above, an IE procedure with a domain-specific ontology was developed based on OSHA Regulations, and then again tested on Construction Workplace Design Solutions.

As shown in Figure 3, concept instances in the ontology were matched with other instances to structure the extracted semantics through a set of relationships among the instances. For example, each building element concept instance (e.g., parapet wall) may have relevant concept instances describing its role and properties. Then, the NLP tool scans the sentences iteratively to generate relationships between concept instances according to the semantic rules defined by the ontology. In the first scanning, the sentences are scanned to look for property concept instances and their associated values. Each property concept instance is associated with a value, that is, a cardinal number that may be followed by a measurement unit concept instance by the order of its appearance in the sentence. ñHighò and ñstrongò are property concepts in the above example, which are attached to has_quantity slots, such as ñ39 inchesò and ñ200 pounds,ò respectively. In the following sentences, the verb ñfunctionò is checked together with the adjacent concept instance, ñbarrier.òThis may represent the relationship of the has_function.

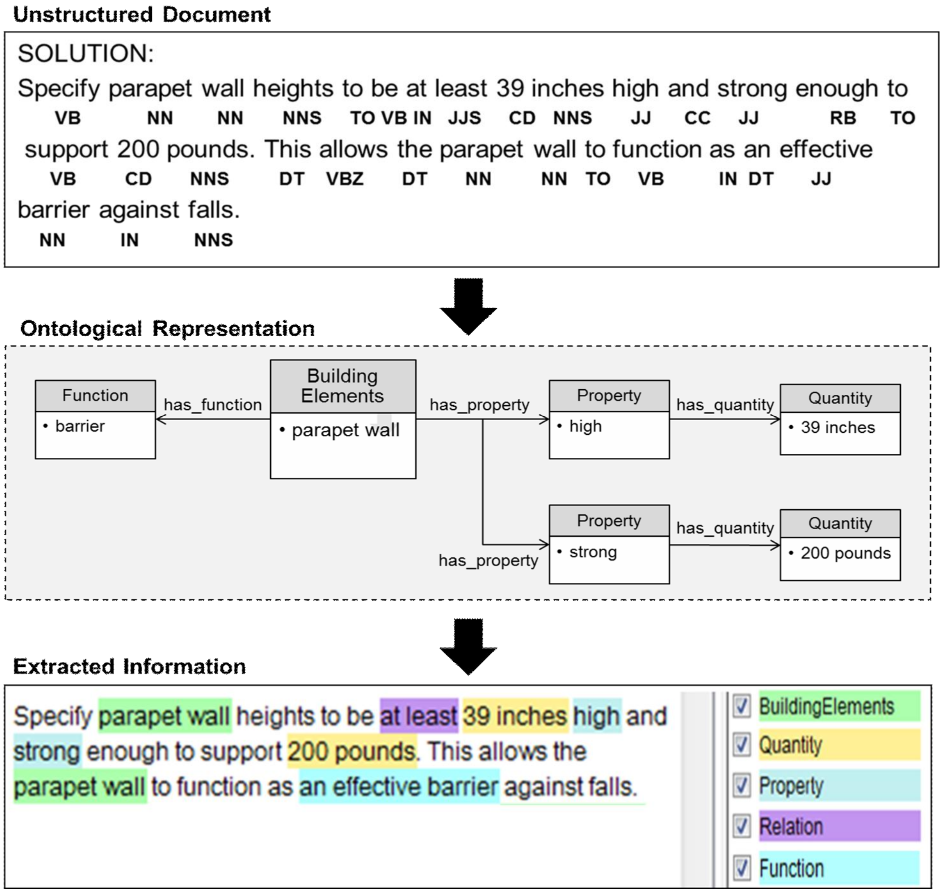

Unstructured Document
SOLUTION:
Specify parapet wall heights to be at least 39 inches high and strong enough to
VB NN NN NNS TO VB IN JJS CD NNS JJ CC JJ RB TO
support 200 pounds. This allows the parapet wall to function as an effective
VB CD NNS DT VBZ DT NN NN TO VB IN DT JJ
barrier against falls.
NN IN NNS
Ontological Representation
Function
• barrier
has_function
Building Elements
• parapet wall
has_property
Property
• high
has_quantity
Quantity
• 39 inches
has_property
Property
• strong
has_quantity
Quantity
• 200 pounds
Extracted Information
Specify parapet wall heights to be at least 39 inches high and strong enough to support 200 pounds. This allows the parapet wall to function as an effective barrier against falls.
☑ BuildingElements
☑ Quantity
☑ Property
☑ Relation
☑ Function

Figure 3: Example of semantic information extraction

Preliminary tests were conducted to evaluate the effectiveness of the IE procedure using ontology, in comparison to the IE without ontology, in improving the performance of retrieving the pertinent information. The preliminary results are summarized in Table 2.

Table 2: Comparison of information extraction results

	IE without Ontology	IE with Ontology
# of correct instances	30	30
# of instances extracted	36	28
# of instances correctly extracted	12	26
Precision	33.3%	92.9%
Recall	40.0%	86.7%
F-measure	36.3%	89.7%

In the case of the construction workplace design solutions, there were 30 tuples that we should extract in the sample document. After the IE without ontology, we extracted 12 correct instances (recall is 40.0%) out of 36 extracted instances (precision is 33.3%). However, ontology-based IE generated 26 correctly extracted instances (recall is 86.7%) out of 28 extracted instances (precision is 92.9%) in total. This result indicates that the proposed ontology-based IE approach is effective in extracting hazard information from an unstructured textual document. It should be noted that these accuracy measures are used only for the task of identifying instances of property values. Evaluating the quality of a used ontology is quite subjective and beyond the scope of this study.

4. Conclusions and Recommendations

As part of the effort to fully automate conformance checking, we presented an automated procedure for extracting safety regulatory information using natural language processing techniques and ontology by expanding our previous effort (Kwon et al. 2013). An ontology-based IE approach was proposed to overcome the matching of lists and single patterns. The proposed IE procedure was applied to the two different documents for construction worker safety that are composed of different words in different sentence structures. The experimental results showed that our approach effectively extracts safety regulatory information. The proposed IE procedure is expected as a basis for fully automated conformance checking.

The remaining issues will be addressed in future works. In this study, the IE procedure was applied to only two types of documents: OSHA Regulations and Construction Workplace Design Solutions. The IE procedure will be applied to other documents to extend its applicability and usability. Additionally, an effective way of manipulating a domain-specific ontology will be considered. These future studies are required to bring this study to its full application in the construction industry, and are expected to promote the development of automated conformance checking tools.

5. References

Behm, M. 2005. Linking construction fatalities to the design for construction safety concept, *Safety Science*, 43(8): 589ï 611.

Bureau of Labor Statistics. 2012. *National Census of Fatal Occupational Injuries in 2011* (Preliminary Results), United States Department of Labor, Washington, D.C.

Eastman, C.M., Lee, J., Jeong, Y., Lee, J. 2009. Automatic rule-based checking of building designs, *Automation in Construction*, 18(8): 1011ï 1033.

Gambatese, J.A. 2008. Research issues in prevention through design, *Journal of Safety Research*, 39: 153ï 156.

Gambatese, J.A., Hinze, J., and Behm, M. 2005. Investigation of the viability of designing for safety, *The Center to Protect Workers' Right*, MD.

Gambatese, J.A., Hinze, J.W., and Haas, C.T. 1997. Tool to design for construction worker safety, *Journal of Architectural Engineering*, 3(1): 32ï 41.

Gibb, A., Haslam, R., Hide, S., and Gyi, D. 2004. The role of design in accident causality, *Proceedings from a Research and Practice Symposium*, September 15ï 16, Portland, OR, USA, 11ï 21.

Grishman, R. 1997. *Information Extraction: Techniques and Challenges*, in: M. Pazienza (Ed.), Springer, Berlin, Germany.

Holt, A.J. 2001. *Principles of Construction Safety*, 1st Ed., Blackwell Science Ltd., Oxford, UK.

Howard, J. 2008. Prevention through designï Introduction, *Journal of Safety Research*, 39: 113ï 113.

Jurafsky, D., and Martin, J.H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd Ed., Prentice-Hall, Upper Saddle River, New Jersey.

Kwon, J., Kim, B., Lee, S., and Kim, H. 2013. Automated hazard identification framework for the proactive consideration of construction safety, *The 5th International Conference on Construction Engineering and Project*, January 9ï 11, CA, USA.

Maynard, D., Funk, A. and Peters W. 2009. SPRAT: a tool for automatic semantic pattern-based ontology population. *In International Conference for Digital Libraries and the Semantic Web*, Trento, Italy, September.

Szymberski, R. 1997. Construction Project Safety Planning. *TAPPI Journal*, 80(11): 69ï 74.

Waehrer, G.M., Dong, X.S., Miller, T., Haile, E., and Men, Y. 2007. Costs of occupational injuries in construction in the United States, *Accident Analysis and Prevention*, 39(6): 1258ï 1266.

Zhou, W., Whyte, J., and Sacks, R. 2012. Construction safety and digital design: a review, *Automation in Construction*, 22: 102ï 111.