Analysis of Construction Accidents Based on Semantic Search and Natural Language Processing

Seonghyeon Moon[a], Taekhyung Kim[a], Bon-Gang Hwang[b], and Seokho Chi[c]

[a]Graduate Student, Department of Civil and Environmental Engineering, Seoul National University, South Korea
[b]Associate Professor, Department of Building, National University of Singapore, Singapore
[c]Associate Professor, Department of Civil and Environmental Engineering, Seoul National University, South Korea
E-mail: blank54@snu.ac.kr, slelic@snu.ac.kr, bdghbg@nus.edu.sg, shchi@snu.ac.kr

Abstract –

Retrieving proper accident cases and extracting risk factors from them are crucial for construction safety management. However, the process was often challenging due to unstructured properties of text data in accident reports, which caused limited, inefficient, and non-consistent information retrieval and knowledge gathering. To overcome the problems, this research aimed at developing a semantic search system to retrieve proper accident cases based on user's deliberate intentions and to extract safety risk factors automatically using Natural Language Processing. The performance of the system prototype was evaluated by construction practitioners with promising results for more usable construction accident database development (i.e., thesaurus) and efficient accident analysis using the thesaurus.

Keywords –

Construction Accident, Safety Management, Semantic Search, Natural Language Processing

1 Introduction

The construction industry is known to be one of the most dangerous industries. According to the Korea Occupational Safety & Health Agency (KOSHA), 29.3% of occupational accident occurrences (26,570 records) were from construction sites in 2016. Moreover, the number of victims by the construction accidents has been steadily increasing for the last three years [1].

The Korean Society of Civil Engineers (KSCE) published a research report of which contents insisted that construction accidents showed similar patterns in occurrence. The patterns consist of (1) what caused the accident, (2) where it occurred, (3) when it was, and (4) how the results came out. These attributes are referred to construction accident risk factors one by one: (1) hazard object, (2) hazard position, (3) work process, and (4) accident result [2].

Since the construction accidents occur in similar patterns, many studies determined that it is possible to prevent construction accidents by identifying and eliminating the risk factors by analyzing similar cases [3,4,5,6,7]. For this reason, retrieving proper cases and extracting risk factors are crucial to preventing construction accidents.

To support academic and industrial efforts on accident analysis, two public institutions operate and manage construction accident databases. One is KOSHA, a database run by itself, and another is Construction Safety Management Information System (COSMIS), run by Korea Infrastructure Safety Corporation (KISTEC). The integrated data of the two databases cover more than 4,000 records of historical construction accidents.

However, the construction accident cases are written in the form of text document which is difficult to be managed and analyzed [8]. The situation generates some troubles in searching and analyzing accidents. The search algorithm implemented on the current database follows binary search "same or different", hence it could only retrieve results exactly matching the given query word by word precisely. This might limit useful information of previous cases, especially due to many synonyms and natural languages used in the construction domain, which would be used as basis for conducting construction accident analysis [9,10]. Besides, accident analysis process would be inefficient and the results could have non-consistency due to laborious task such as perusing numerous accident reports and labeling risk factors manually [11,12,13].

Taken altogether, analyzing construction accident cases based on current databases is difficult due to the unstructured properties of text documents. To overcome these problems, the research aimed at developing a semantic search system supporting two major functions: to retrieve proper cases based on user's deliberate intentions and to analyze accident cases by extracting risk factors automatically.

Accessing KOSHA and COSMIS, the research

collected 4,263 reports of construction accidents which had occurred from September 1 1990 to October 18 2017. Natural Language Processing (NLP) was used to manipulate the text data of the reports and conduct the research. Meanwhile, the research adapted pre-existing risk categories identified by the Construction Risk Factor Profile [2] in order to define the labels of construction accident risk factor as mentioned in the first paragraph.

2 Research Methodology

2.1 Web Crawling

Construction accident reports were collected from KOSHA and COSMIS websites by the web crawling method which is a web-based process that accesses websites and collects target data [14]. Websites are usually constructed with Hypertext Markup Language (HTML), a standard language of websites used to specify detail features of entries such as position, font, color, and size. Once the web crawler is set to extract certain information, it finds the tags (e.g., <position>, <font>, <color>, <size>, etc.) and takes the information from each tag.

The process consists of two steps: (1) list page parsing and (2) target page parsing [15]. In the first step of list page parsing, the web crawler (i.e., web crawling algorithm) extracts Uniform Resource Locator (URL) links of target pages from the list page that covers the user's query. Afterwards, during target page parsing, the web crawler extracts actual data from the target pages. The procedure is provided in Figure 1.

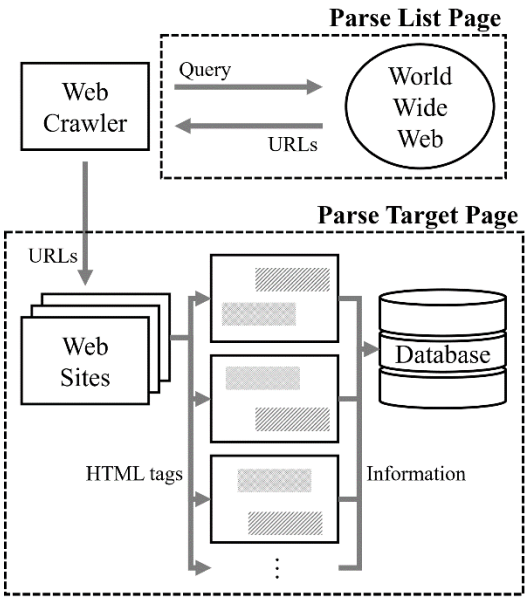

Figure 1 Web Crawling Procedure

2.2 Preprocessing

Text data preprocessing was conducted in order to convert raw data into an understandable format for the computer by splitting text strings into several valid words [15]. In this research, the preprocessing was composed of four steps. First was data cleaning; all punctuation marks were eliminated since those non-verbal items were worthless to understanding the text data. Second was tokenizing; every sentence from accident reports was split into a single word (i.e., token) based on space marks. Third was normalizing; each token was converted into its canonical form. Fourth and the last was stopword removal. Stopwords (i.e., less informative tokens) including grammatical components and person's name were removed to improve data quality. For instance, a sentence "During installation process, John lost his balance and fell down." would be converted into "installation process lose balance fall down" after preprocessing as explained in Figure 2.

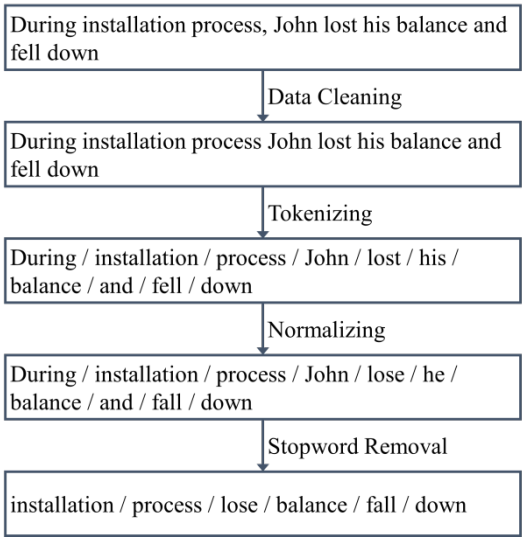

Figure 2 Procedure of Text Preprocessing

2.3 Query Expansion

For the purpose of including the user's deliberate intention to queries, the newly developed search system is able to expand the given queries using the thesaurus [16]. According to ISO 25964-1 standard, a thesaurus is a kind of advanced dictionary containing additional information such as the synonym, hypernym, and hyponym of each term [17]. Generally, plenty of advanced search engines already have been adopting query expansion using a thesaurus to overcome uncongenial result problems [18,19]. The research established a construction accident thesaurus using several construction dictionaries and a list of similar terms derived from the Word2Vec algorithm.

Specifically, twelve construction dictionaries with the total of 15,564 terminologies were adopted to the system. The construction dictionaries represented term relationships in the construction industry (e.g., synonym, hypernym, hyponym.), hence query expansion using dictionaries could be regarded as considering the relationship of meanings between terms.

Word2Vec is an algorithm which creates a corresponding vector for each term based on distribution similarity of surrounding terms [20]. For instance, in the two sentences "A worker got hurt on head" and "A worker got injured on hand", the term "hurt" and "injured" will be mapped close to each other since the surrounding terms are very similar. Consequently, it could be deducted that the Word2Vec algorithm clustered terms using the similarity of term usages which existed in the accident reports.

2.4 Frequent-based Search

The last steps to the search algorithm include assessing relevance of documents and returning the search results. To get a quantitative achievement in relevance assessment of documents, the research adopted a frequency based ranking method called BM25 which considers two frequency indicators – Term Frequency (TF) and Inverse Document Frequency (IDF). TF is the number of occurrences of a term in a document, which implicates the importance of a term. That is, the larger the TF is, the more important the term is. However, if a term exists in every document, then the term cannot be used as a distinctive feature for judging the importance among documents. For this reason, IDF, an inverse number of documents which contain a certain term, was designed to normalize the scale effect of TF [15,21]. Since higher BM25 score means the document is more relevant to the query, the accident reports are sorted in descending order.

After identifying user-wanted accident reports based on the BM25 score, predetermined four kinds of risk factors were extracted by conducting Named Entity Recognition (NER). NER is a text classification method which labels each term with informative categories such as person's name, place, object, action, and so on [22]. Among existing NER algorithms, the research employed Conditional Random Field (CRF) to determine each term's label on the basis of conditional probability of the surrounding terms, hence had been showing the highest performance at text labelling [23]. The conditional probability of each class y for a target token is calculated as Equation (1), where Y represents a set of labels (i.e., set of risk factors) X represents a sequence of observations which equals to the surrounding terms. On the right side of the Equation (1), t_j and s_k represents feature functions of which output is 0 or 1, λ_j and μ_k are weighting values for each feature function, and Z(X) is a constant normalizing the maximum probability as 1. Consequently, the target token was labelled by the class with the largest probability as provided in Equation (2), where y^* indicates the label of the target. The correct dataset was constructed by 51 rules which labelled terms related to risk factors from each sentence base on the relationship between sentence elements [24].

$$P_\Lambda(Y|X) = \frac{1}{Z(X)}\exp\left(\sum_j \lambda_j t_j + \sum_k \mu_k s_k\right) \quad (1)$$

$$y^* = \underset{Y}{argmax}\, P_\Lambda(Y|X) \quad (2)$$

3 System Prototype

A system prototype was developed to test the proposed methodology and evaluate the results. Figure 3 provided the result page of the system prototype with the query "타워 크레인 추락" which means "Tower

Crane Fall" in English. The system prototype was developed by localhost, which means only local computers have authority to access. This is to block unwanted users, since the research is in experimental stage.

When a user gave a query "tower crane fall" as an input, the system preprocess the query with data cleaning, tokenizing, normalizing, and stopword removal. The query would be divided into two tokens "tower crane" and "fall", then expanded based on the thesaurus to various queries: "tower crane", "crane", "lifting equipment", "lift", "T/C", "coping", "jib", "winch", "falling", "drop", "collision", "fall beneath", etc. The expanded queries were visualized in the top of the prototype as the red box in Figure 3. One of the most powerful text visualization tool, Word Cloud, was applied to show the queries, of which the font size indicates importance and leverage [25,26].

After searching and ranking appropriate documents with the queries based on the BM25 score, the system retrieved the most proper accident reports right below the Word Cloud. The accident name and short description of each case are provided in rows as marked with the dark blue box in Figure 3. For detailed information, users can access the original report simply by clicking the case.

Eventually, the CRF model extracted risk factors from the text data of retrieved reports. The risk factors (i.e., hazard object, hazard position, work process, and accident result) were provided on the right side of accident descriptions, marked in orange, light blue, green, and purple boxes each. For instance, the first case was "Haeundae I-park tower crane collapsed" of which content was that "Jib of tower crane collapsed due to sudden gust". The risk factors of this case were identified as "jib" for the hazard object and "collapse" for the accident result. As some of accident reports did not specify every risk factor explicitly, several blanks existed in the result.

Figure 3. Layout of System Prototype

4 Validation

4.1 Validation of Query Expansion

The retrieval results based on query expansion were validated in qualitative approach using Normalizing Discounted Cumulative Gain (NDCG), a commonly used approach to evaluate information retrieval results. NDCG examined whether relative documents were ranked high in the retrieval result by comparing documents from the result and correct set at every rank position p [27]. The NDCG is defined as Equation (3), where rel_i is the relevance of the result at position i and $|REL|$ represents the number of relevant documents up to position p [28].

$$NDCG_p = \frac{\sum_{i=1}^{p} 2^{rel_i} - 1 / \log_2(1+i)}{\sum_{i=1}^{|REL|} 2^{rel_i} - 1 / \log_2(1+i)} \quad (3)$$

To secure testing set, one query and 10 documents were given to each of 4 questionnaire groups composed of 4 construction practitioners with less than 5-year work experience. Then, each group ranked the documents based on the relativity of documents with the given query with 5-point scale: '5' represents the very high level relevance and '1' represents the very low level. NDCG score was calculated using the ranked results for four queries. Finally, the research compared the retrieval results with the testing set which questionnaires ranked and quantified the difference based on NDCG. The scores of each query were 0.98, 0.98, 0.94, 0.98, where 1.00 represents a consensus and otherwise 0.

4.2 Validation of Risk Factor Extraction

In the first step, the relevance of correct set generated from rule-based approach was verified. Same questionnaires with the former step were involved to label every risk factor from 101 documents. Then, the results of rule-based labeling were compared to the results of human-based. As a result, the rule-based approach showed an average accuracy of 93.75% establishing its relevance as training set for the CRF model. Specifically, labeling of hazard object scored 96% (82 from 85), hazard position 87% (46 from 53), work process 95% (82 from 86), and accident result 97% (98 from 101) which are provided in Table 1. The population of each risk factor would be different since not every accident report contained all of the risk factors. Since the relevance scores were considerably high, the CRF model, trained by the rule-based labeled data, must be trained attentively.

Table 1 Verification of Rule-based Labeling

Risk Factor	Relevance of Rule-based Labeling	
Hazard Object	82 / 85	96 %
Hazard Position	46 / 53	87 %
Work Process	82 / 86	95 %
Accident Result	98 / 101	97 %

Next, the labeling results of the CRF model were validated using the remaining 10% of data based on precision and recall of each risk factor. As provided in Table 2, most of the precision and recall scores came out to be sufficient to insist the feasibility of the system. Even though the recall rate of the hazard position was evaluated quite low, the overall results showed promising applicability of the proposed system. The precision and recall rates would become higher as the accident reports are accumulated constantly.

Table 2 Validation of CRF Model

Risk Factor	Precision	Recall
Hazard Object	0.95	0.68
Hazard Position	0.93	0.52
Work Process	0.86	0.74
Accident Result	0.99	0.90

5 Conclusions

Analyzing construction accident cases is crucial for safety management, since construction accidents occur in similar patterns which could be represented by hazard objects, hazard positions, work processes, and accident results. However, the current databases collecting accident reports have limitations to retrieve and analyze accident cases due to unstructured properties of text reports. This research thus developed a prototype of semantic search systems that retrieve similar cases based on the user's deliberate intentions and analyze accident cases by extracting risk factors automatically by applying text-mining techniques.

The developed system would have several contributions to both of industrial and academic sectors. For the industry level, the system would help provide accident references for safety managers to explain the causes of accidents and precautions to prevent them by applying the developed thesaurus and NER. For the academic level, the research identified relationships between keywords that exist on construction accident reports by using the Word2Vec algorithm and eventually explained the feasibility of NLP-based text analyses for better construction safety management.

6 ACKNOWLEDGMENT

This research was supported by a grant(16CTAP-C114956-01) from Technology Advancement Research Program (TARP) Program funded by Ministry of Land, Infrastructure and Transport of Korean government and Seoul National University Big Data Institute through the Data Science Research Project 2017.

References

[1] KOSHA. Industrial Accident Occurrence in 2016. On-line: http://www.kosha.or.kr/board.do?menuId=554, Aceessed: 19/01/2018 (Korean)

[2] KSCE. (2014). Research Report of Developing Construction Risk Factor Profile. (Korean)

[3] Baradan S. and Usmen M. A. Comparative injury and fatality risk analysis of building trades. *Journal of Construction Management in Engineering*, 132(5):533–539, 2006.

[4] Hallowell M. and Gambatese J.A. Activity-based safety risk quantification for concrete form work construction. *Journal of Construction Management in Engineering*, 135:990–998, 2009.

[5] Park M., Lee K. W., Lee H. S., Pan J. Y., and Yu J. Ontology-based construction knowledge retrieval system. *KSCE Journal of Civil Engineering*, 17(7):1654–1663, 2013.

[6] Shapira A. and Lyachin B. Identification and analysis of factors affecting safety on construction sites with tower cranes. *Journal of Construction Management in Engineering*, 1351:24–33, 2009.

[7] Tixier A. J. P., Hallowell M. R., Rajagopalan B., and Bowman D. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62:45–56, 2016.

[8] Soibelman L., Wu J., Caldas C., Brilakis I., and

Lin K. Y. Management and analysis of unstructured construction data types. *Advanced Engineering Informatics*, 22(1):15–27, 2008.

[9] Holscher C. and Strube G. Web search behavior of internet experts and newbies. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 33(1–6):337–346, 2000.

[10] Spink A., Wolfram D., Jansen B. J., and Saracevic T. Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 53(2):226–234, 2001.

[11] Desvignes M. Requisite empirical risk data for integration of safety with advanced technologies and intelligent systems. *Master's Thesis, University of Colorado, Boulder*, 2014.

[12] Esmaeili B. and Hallowell M. Attribute-based risk model for measuring safety risk of struck-by accidents. In *Proceedings of the Construction Research Congress 2012*, pages 289–298, 2012.

[13] Prades M. Attribute-Based Risk Model for Assessing Risk to Industrial Construction tasks. *Master's thesis, University of Colorado, Boulder*, 2014.

[14] Cho J. Crawling the web: discovery and maintenance of large-scale web data. *Ph. D. Thesis, Stanford University, Stanford, CA*, 2002.

[15] Manning C. D., Raghavan P., and Schutze H. *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[16] Vechtomova O. and Wang Y. A study of the effect of term proximity on query expansion. *Journal of Information Science*, 32(4):324–333, 2006.

[17] ISO 25964-1: *Information and documentation - Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval*, 2011.

[18] Colace F., De Santo M., Greco L., and Napoletano P. Weighted word pairs for query expansion. *Information Processing & Management*, 51(1):179–193, 2015.

[19] Gao, G. Liu Y. S., Wang M., Gu M., and Yong J. H. A query expansion method for retrieving online BIM resources based on Industry Foundation Classes. *Automation in Construction*, 56:14–25, 2015.

[20] Mikolov T., Sutskever I., Chen K., Corrado G. S., and Dean J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in neural information processing systems*, pages 3111–3119, 2013.

[21] Robertson S. and Zaragoza H. The probabilistic relevance framework: BM25 and beyond. *Foundation and Trends in Information Retrieval*, 3(4):333–389, 2009.

[22] Nadeau D. and Sekine S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1): 3–26, 2007.

[23] Lafferty, J. McCallum A., and Pereira F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, 2001.

[24] Yoo H. W. The study of the methodology of the Korean parser. *Korean Culture Research*, 50(0):153–182, 2009.

[25] Cui W., Wu Y., Liu S., Wei F., Zhou M., and Qu H. Context Preserving Dynamic Word Cloud Visualization. In *Proceedings of the IEEE Pacific Visualization Symposium 2010*, pages 121–128, Taipei, Taiwan, 2010.

[26] Heimerl F., Lohmann S., Lange S., and Ertl T. Word Cloud Explorer: Text Analytics based on Word Clouds. In *Proceedings of the* 47th *Hawaii International Conference on System Science*, 1833–1842, 2014.

[27] Järvelin K. and Kekäläinen J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422-446, 2002

[28] Wang Y., Wang L., Li Y., He D., Chen W., and Liu T. Y. A theoretical analysis of Normalized Discounted Cumulative Gain (NDCG) ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory* (COLT 2013), 2013.