NLP and Deep Learning-based Analysis of Building Regulations to support Automated Rule Checking System

Jaeyeol Song[a], Jinsung Kim[a], and Jin-Kook Lee[a]

[a]Department of Interior Architecture & Built Environment, Yonsei University, Republic of Korea
E-mail: songjy92@gmail.com, wlstjd1320@gmail.com, leejinkook@yonsei.ac.kr

Abstract –

This paper aims to describe a natural language processing (NLP) and deep learning-based approach for supporting automated rule checking system. Automated rule checking has been developed in various ways and enhanced the efficiency of building design review process. Converting human-readable building regulations to computer-readable format is, however, still time-consuming and error-prone due to the nature of human languages. Several domain-independent efforts have been made for NLP, and this paper focuses on how computers can be able to understand semantic meaning of building regulations to intelligently automate rule interpretation process. This paper proposes a semantic analysis process of regulatory sentences and its utilization for rule checking system. The proposed process is composed of following steps: 1) learning semantics of words and sentences, 2) utilization of semantic analysis. For semantic analysis, we use word embedding technique which converts meaning of words in numerical values. By using those values, computers can extract related words and classify the topic of sentences. The results of the semantic analysis can elaborate the interpretation with domain-specific knowledge. This paper also shows a demonstration of the proposed approach.

Keywords –

Automated rule checking; Natural language processing; Deep learning; Semantic analysis; Building information modeling (BIM)

1 Introduction

As BIM (Building information modeling) has increasingly used in architecture, engineering, construction (AEC) industry, there are many efforts to develop BIM applications to use rich building information. The role of automated rule checking has been recognized before the use of BIM [1]. As BIM provides computer interpretable building model, many researchers reported that BIM can improve efficiency and accuracy of design assessment [2]. The main components of rule checking consist of 1) rule interpretation and logical structuring, 2) building model preparation, 3) rule execution, and 4) reporting of the checking results [3]. Compared to other phases, rule interpretation task needs a lot of manual efforts. For automated rule checking, regulatory information must be expressed in an explicit format which computers can execute. Extracting and structuring required data from sentence is time-consuming and error-prone due to the vagueness of natural language. Therefore, to intelligently automate rule interpretation, computers should understand the semantics of the regulatory sentence.

Deep learning-based NLP (Natural language processing) enables the computers to learn the semantics of natural language from raw text data. Unlike conventional method, deep learning model extracts training features from raw text and adjusts the weight without human's intervention. It can alleviate the manual efforts to interpret regulation sentences which required for automated rule checking. In this regards, the objective of this research is to apply deep learning-based NLP for translating building regulations into a computer-readable format. As an early stage of research, this paper proposes a semantic analysis process to support rule checking, focusing on rule interpretation.

The scope of this paper is the approach to analyzing the semantic data of Korean building regulation sentences with NLP and deep learning. We analyzed 5 Korean building regulations: Building Act, Enforcement Decree of Building Act, Regulation of the building Structure criteria, Regulation of the building facility criteria, and Regulation of the evacuation and fireproof construction criteria. 3504 training sentences are derived from 5 building regulations. Steps of the proposed analysis are summarized as follows.

1. Learning step: Decompose training sentences in word-level and learn semantic of words and sentences by using NLP and deep learning model
2. Utilization step: Automatically inference the related word and topic of input sentence, utilizing trained learning model.

After dealing with proposed analysis process, this paper also presents a demonstration of utilization which is associated with domain database.

2 Background

2.1 Rule interpretation for automated rule checking

There have been many efforts to develop automated rule checking. Among the automated rule checking process, rule interpretation is a significant step for automating rule checking. The conventional methodology to translate the building code has been depended on hardcoding or logic rule-based mechanism. The CORENET (COnstruction and Real Estate NETwork) project in Singapore translated building code of Singapore into the programming language by hardcoding [4]. DesignCheck project formalized the building code of Australia and encoded it according to the EDM rule schema based on the EXPRESS language [5]. In GSA project, the courthouse design guide was parameterized and translated into computer-processible format [3]. In case of the Korean building permit research project, they used logic rules and intermediate code to translate Korean building act into the computer-readable format [6]. The conventional rule-based way guarantees the accuracy and reusability. However, establishing the logic rule for interpretation has been done by manual efforts due to ambiguity and vagueness of natural language.

In order to overcome the inefficiency of a manual process, some studies focused on analyzing the semantics of words in regulatory sentences and utilizing NLP. Zhang et al. proposed semantic NLP-based automated compliance checking (SNACC) system [7]. NLP and ontology were applied for extracting the regulatory information from documents. Through the NLP technique and other process, they automated the rule interpretation with semantic information extraction. E.Hjelseth applied RASE methodology to transform normative documents into a single well-defined rule which can be implemented into model checking software [8]. Uhm et al. translated request for proposal (RFP) for public building in South Korea [9]. Sentences in RFP were broken down into morphemes and categorized into four types (object, method, strictness, and others). Context-free grammar approach was deployed for parsing, and the sentences were translated into Semantic Web Rule Language (SWRL). SWRL is translated again into Python script language for implementation. Previous researchers have contributed to capturing the semantics of regulatory documents, but they still need manual effort to interpret patterns of sentences.

2.2 Natural language processing

NLP is a research field of artificial intelligence related to interactions between human and computer. The main research of NLP ranges from understanding human language to making proper responses. Due to the ambiguity of natural languages, it is a challenging process to understand the raw text data. Conventional NLP used logic rule-based or statistical methodology for analyzing natural language. In the rule-based method, the text is analyzed by rules defined with linguistic knowledge. However, the semantic meaning of words is hard to capture since it is hard to be defined by pattern or rules. WordNet is a lexical database for English, which provides sets of synonyms that are in turn linked through semantic relations [10]. This has been widely utilized to present semantic meaning of words but still needs a manual task to build and manage this database.

Development of machine learning and deep learning has changed the methodology of NLP and dramatically improved the accuracy of it [11]. Deploying an extensive amount of data and neural net algorithm, the computer can learn semantics in natural language by itself. As machine learning is processed with numerical vectors, words have to be converted into the computer-interpretable format. Distributed representation in vector space helps to represent words quantitatively. Rumelhart proposed this concept of analysis [12] and recently Mikolov implemented neural net based word embedding model called word2vec [13, 14]. After google opened the word2vec model to the public, many researchers make use of this model for sentiment analysis [15] and information extraction [16].

3 Learning semantic of words and sentences in building regulations

This section describes details of the semantic analysis for building regulations. As shown in Figure 1, proposed process is composed of 2 main steps: Learning semantics and Utilization. Learning step consists of 1) preprocessing, 2) semantic analysis of words and sentences and 3) sentence classification. The process and details about utilization are described in Section 4.

Preprocessing is focused on the grammatical analysis of documents, which is needed for semantic analysis. After the preprocessing, the meaning of words and the topic of sentences are learned by neural net-based word embedding technique. Sentence classification is conducted with deep learning model. In this paper, we used a Python package library KoNLPy [17] for preprocessing, Gensim [18] for word embedding and Tensorflow [19] for establishing classification model.

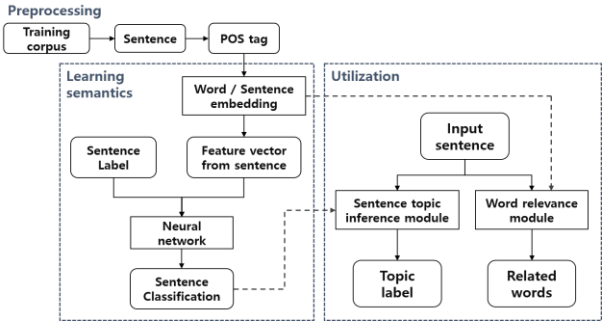

Figure 1. Proposed process of analysis and utilization of building regulations

3.1 Preprocessing

In order to proceed an accurate semantic analysis, raw text data have to be processed in a suitable format. In the Korean language, there are postposition particles which stand for grammatical use. They are attached behind to other words like nouns or adjectives. In the raw text, there are also stop words which disrupt semantic analysis (e.g. punctuation). The sentences should be analyzed in grammatically to exclude these unnecessary part. Preprocessing consists of morpheme analysis, Part-of-Speech (POS) tagging and excluding stop words. Morpheme is the smallest unit which has a specific meaning in a language. By decomposing words in atomic unit, morpheme analysis supports to separate the words from postposition particle. POS tagging assigned the label of grammatical function to each word, such as noun, verb, and adjectives. After POS tagging, we excluded Chinese character, punctuations, and numbers. Figure 2 presents a result of each step of preprocessing.

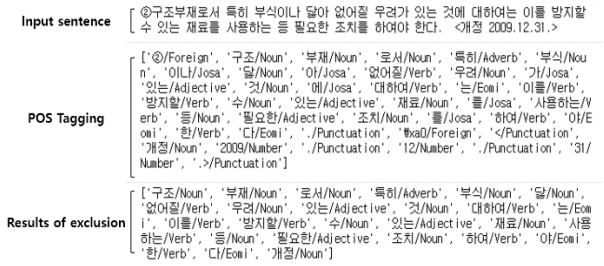

Figure 2. Example of preprocessing result (POS-tagging, Exclude stop words)

3.2 Semantic analysis of words and sentences

This paper uses word2vec model for the semantic analysis of words. Word2vec model is based on the word embedding technique and it learns the semantics of text from co-occurrence information. Some words which appear with target word in the same context are more related to target word than the others. Based on this concept, Mikolov et al. proposed 2 models to learn the word vector 1) CBOW (continuous Bag-of-Words) model and 2) Skip-gram model [14]. Skip-gram model predicts the current word with a certain range before and after the current word, CBOW model is the opposite. By predicting the current or context words, the computer can learn a co-occurrence data of words. Based on the co-occurrence data, the model assigns the high-dimensional vector values for each word. The numerical vector values enable for the computer to calculate a relevance of words. This learning process also can be applied to learning the meaning of sentence [20]. In this paper, the meaning of words and sentences are presented in 400-dimensional vector space, as shown in Figure 3. We set the maximum distance 5 and words with a frequency less than 3 were excluded from the training.

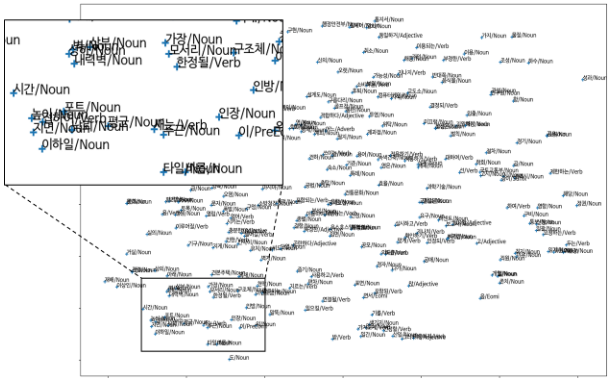

Figure 3. t-SNE Visualization of word distribution

3.3 Sentence classification

Korean building regulations contain not only requirements for building also other rules which are irrelevant to buildings (e.g. administrative procedures, Committees). BIM-based automated rule checking focuses on the requirements for buildings which can be expressed quantitatively. Therefore, computer should classify the sentences according to the content and extract the sentences related to buildings from the raw documents. Furthermore, classifying sentences supports identifying what information has to be extracted. The numerical vector values extracted from semantic analysis facilitate the classification of the sentences.

In this paper, we use deep learning model for classification of sentences. First, we tag the label to each sentence according to its contents. Table 1 shows a classification of content described in building regulations. In this paper, 6 categories are used for classification labels: 1) Non-AEC, 2) Site, 3) Building structure, 4) Facility, 5) Usage of building and 6) Evacuation & Fireproof. We just use Non-AEC category, as we don't need more detailed for irrelevant sentences in terms of

rule checking. Site category covers the regulations for a building site and roads. Facility category includes MEP facilities and Evacuation & Fireproof category is comprehensive of rules for evacuation plan and noncombustible material. For supervised classification learning, we use deep learning model which composed of 4 fully-connected layers and softmax function for multi-class classification. The graph of model and weights are saved for inference model.

Table 1. Classification of contents described in building regulations

Level-1	Level-2	Level-3
Non-AEC	Rule	Amendment
		Delegation, Reference
	Administrative	Procedure
		Committees
		participant
AEC	Site	Site
	Building	Structure
		Facility
		Usage
		Evacuation & Fireproof

4 Utilizing semantic analysis to support automated rule checking system

4.1 Extract related words from input sentence

Extracting related words to target word facilitates understanding of the semantics. Based on the numerical values, the computer can calculate the metric distance between words. The distance between words represent the relevance between them, so we can extract words related to the target word. This features could be used for extract semantic relations of words in a regulatory sentence.

This module extracts words related to a user-input word, using results of the semantic analysis. The given sentence which is selected by users is decomposed to word-level. After decomposition, the relevance of input word and each word from the given sentence is calculated. Through those steps, the computer can get a list of related words, but it has no idea which words are building object or their property. This gap can be decreased by domain knowledge which deals with rule checking data. This paper used the KBimCode database made up of previous work of this research [6]. KBimCode database provides the computer-executable code data and corresponding Korean words. It can be used as a dictionary for translating Korean words into object, property and method data, as shown in Figure 4.

Figure 4. Translating words to corresponding object/property data

4.2 Inference the topic of input sentence

Topic inference module is based on a result of classification learning. The input sentence is also decomposed in a set of words. After preprocessing, inference module infers the vector values of input sentence based on other sentences vector values. In this paper, we assigned the meaning of the sentence in 400-dimensional space, so the inference module also returns the same shape of value. Then, the inferred vector is given as an input for the classification model. It uses a same graph and weights for predicting the topic.

5 Demonstration

The proposed modules described in the previous section are implemented in GUI applications for supporting rule checking process. This section demonstrates the results of implementation. With this application, users can get related words, corresponding KBim data and a predicted topic of the input sentence.

Table 2. An example of word extraction for Korean building regulation

Input Sentence	[Enforcement Decree of Building Act, article 35] Direct stairs installed on the fifth or upper floor or the second or lower underground floor pursuant to Article 49 (1) of the Act, shall be installed as fire escape stairs or special escape stairs in accordance with the standards prescribed by Ordinance of the Ministry of Land, Infrastructure and Transport: Provided, That the same shall not apply where main structural parts are made of a fireproof structure or non-combustible materials and either of the following is applicable
Input word	Stair

Top 5 related word	Direct / Noun Escape / Noun Floor / Noun Underground / Noun Special / Noun
Extracted word	Direct stairs, escape stairs, Special escape stairs, underground floor, fifth floor, second floor
KBim data	-Object: Stair, Floor -Properties : Stair.isDirect, Stair.isEscape, Stair.isSpecialEscape -Method : getFloorNumber()

Table 2 shows a result of extracting related word when "stair" is given as an input word. The result is presented in GUI application as shown in Figure 5. The properties of the stair are also easily found in a lexical level search. However, in the lexical level, floors where stairs are installed or related objects are not searched. The demonstration shows that not only the properties of the stair are suggested, also the related object (Floor) and its properties (Underground), from the semantic analysis.

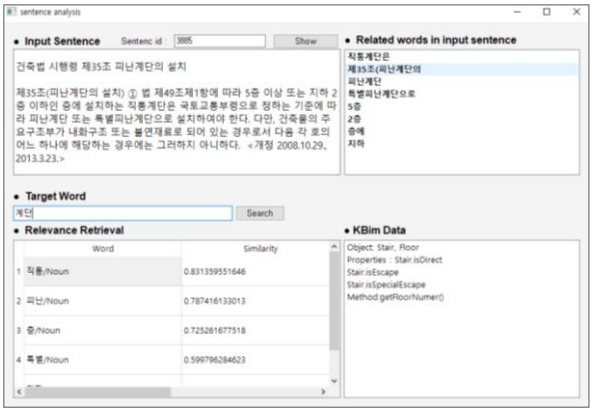

Figure 5. Screenshot of the application for extracting related words in sentence

The topic inference module suggests predicted topic of an input sentence and also similar sentences based on their content. Suggested similar sentences are extracted based on vector representation, calculating the metric distance of each sentence, same process with the word extraction. In terms of extending the scope of rule checking to RFP or other new regulations, this topic prediction makes it easier to classify the sentence whether it deals with content related to BIM-based rule checking. Table 3 shows the examples of topic prediction results. The inference module makes an accurate output in case 1. However, in case 2 and 3 the module make a wrong prediction, as the correct label is predicted in second. The results of prediction are demonstrated in GUI application, as shown in Figure 6.

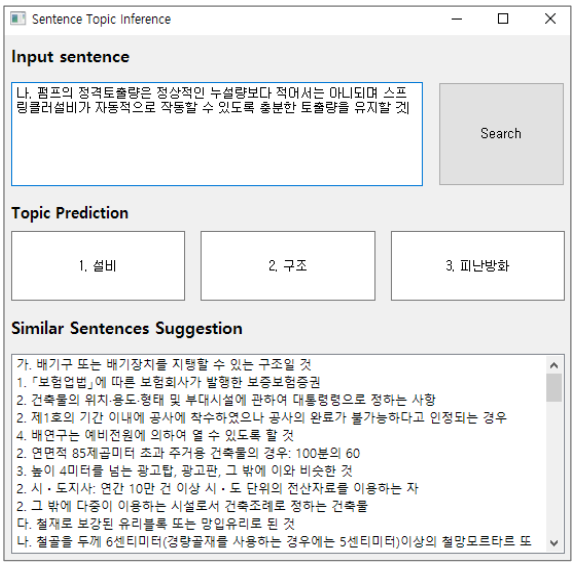

Figure 6. Screenshot of the application for Topic inference module

Table 3. Examples of topic prediction for input sentence

Case	Input Sentence	Output Topic
1	[Fire safety standard for sprinkler (NFSC 103)] Article 5. The rated discharge rate of the pump should not be less than the normal leakage and maintain sufficient discharge volume for the sprinkler system to operate automatically	1. Facility 2. Structure 3. Evacuation &Fire-proof 4. Non-AEC 5. Site 6. Usage
2	[Seoul bylaw for construction] Article 33 Buildings whose floor surface of 1st floor is over 0.5 meters above the ground surface and whose height added to height between 1st floor and ground floor by 8 meters is less than 12 meters.	1. Usage 2. Structure 3. Evacuation & Fire-proof 4. Facility 5. Site 6. Non-AEC
3	[Special law on disaster management of complex buildings and coordination of high-rise and underground] Article 19 Supervisors for high-rise buildings should install and operate shelter safety zone where workers, residents and users can evacuate in the event of a disaster	1. Facility 2. Evacuation & Fire-proof 3. Site 4. Structure 5. Non-AEC 6. Usage

6 Conclusion

This paper proposes a semantic analysis process and its utilization to support rule interpretation. In this paper, Word embedding model is used for learning the semantics of text. Through the training, the semantic of word and sentences are represented in vector values. The results of the semantic analysis are utilized for extracting related words and classifying the topic of sentences. Sentence classification based on deep learning model enables the computer to classify the regulatory sentence according to its content. Extracting related words helps both human and computer to find semantic information from raw text. Demonstrated implementation shows the possibility of utilizing deep learning and NLP to support rule interpretation process. Leveraging this process can decrease the manual efforts and time for interpreting rules.

This paper also has some limitations. In preprocessing process, there are some errors in morpheme analysis which are done by NLP framework. This causes critical problems for information extraction task and interpretation. And this paper mainly deals with the semantic relation of words. To entirely automate rule interpretation, it needs to be developed to the logical relation. For the future work, it can make a breakthrough to deploy the syntax parsing for capture the logical relations of words. Furthermore, to enhance the accuracy of morpheme analysis and sentence inference, other deep learning models could be applied (e.g. RNN or CNN for sentence classification [21]).

Acknowledgement

This research was supported by a grant (18AUDP-B127891-02) from the Architecture & Urban Development Research Program funded by the Ministry of Land, Infrastructure and Transport of the Korean government.

References

[1] Solihin. W., Eastman. C. M., Classification of rules for automated BIM rule checking development, *Automation in construction*, 53: 69-82, 2015

[2] Eastman, C. M., Eastman, C., Teicholz, P., & Sacks, R., *BIM Handbook: A guide to building information modeling for owners, managers, designers, engineers and contractors*, John Wiley & Sons Inc, Hoboken, NJ, United States of America, 2011

[3] Eastman, C. M., Lee, J. M., Jeong, Y. S., Lee, J. K, Automatic rule-based checking of building designs, *Automation in construction*, 18(8):1011-1033, 2009.

[4] CORENET, CORENET Singapore, Available at: http://www.corenet.gove.sg/

[5] Ding, L., Drogemuller, R., Rosenman, M., Marchant, D., Gero, J., Automating code checking for building designs, *Clients Driving Innovation: Moving Ideas into Practice* (pp. 1-16). CRC for Construction Innovation.,2006

[6] Lee, H., Lee, J. K., Park, S., Kim, I., Translating building legislation into a computer-executable format for evaluating building permit requirements. *Automation in Construction*, 71: 49-61, 2016.

[7] Zhang, J., El-Gohary, N. M., Integrating semantic NLP and logical reasoning into a unified system for fully-automated code checking, *Automation in construction*, 73: 45-57, 2017

[8] Eilif Hjelseth, Nick Nisbet, Capturing normative constraints by use of the semantic mark-up RASE methodology, In *proceeding of the CIB W78 28th International conference*, Sophia Antipolis, France 2011.

[9] Uhm, M., Lee, G., Park, Y., Kim, S., Jung, J., Lee, J. K., Requirements for computational rule checking of requests for proposals(RFPs) for building designs in South Korea, Uhm et al., *Advanced Engineering Informatics*, 29(3): 602–615, 2015.

[10] Miller, G. A., WordNet: a lexical database for English, *Communications of the ACM* 38(11): 39-41, 1995

[11] LeCun, Y., Bengio, Y., Hinton, G., Deep learning. *Nature*, 521(7553): 436-444. 2015

[12] Rumelhart, D. E., Hinton, G. E., Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323(6088): 533, 1986

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

[14] Mikolov, T., Chen, K., Corrado, G., Dean, J., Efficient estimation or words representations in vector space, In *ICLR 2013 Workshop paper, Scottsdale, Arizona, USA, 2013*

[15] Xue, B., Fu, C., Shaobin, Z., A study on sentiment computing and classification of Sina Weibo with word2vec, In *Big Data(BigData Congress)*, 2014 IEEE International Congress on, pages 358-363, 2014.

[16] Sienčnik, S. K., Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 239-243, Vilnius, Lithuania, 2015.

[17] Park, E. L., & Cho, S. 2014. KoNLPy: Korean natural language processing in Python. In *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, pages

133-36
[18] Radim Rehurek, Petr Sojka. Software framework for topic modeling with large corpora, *Proceedings of the LREC 2010 Workshop on new challenges for NLP frameworks*, pages 45-50, Valletta, Malta, 2010.
[19] Google, Tensorflow, Online: https://www.tensorflow.org/
[20] Quoc, L., Mikolov, T., Distributed representations of sentences and documents." In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, Beijing, China, 2014.
[21] Kim, Y., Convolutional neural networks for sentence classification, In *Proceedings of the 2014 conference on Empirical Methods in natural language processing*, pages 1746-1751, Doha, Qatar, 2014