Spatial Information Enrichment using NLP-based Classification of Space Objects for School Bldgs. in Korea

J. Song[a], J. Kim[a], and J. Lee[a]

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

Abstract –

This paper presents an approach to classifying spatial categories of space objects using their textual properties in IFC data. As a standardized data format of building information, IFC enhances the data interoperability between the heterogeneous domain software. However, there are some problems that required information is omitted due to the technical translation error or mistake by users. The other problem is that some semantic information cannot be defined in the IFC data scheme. Manually checking and modifying this information requires an amount of time and labor. The intelligent information enrichment system can facilitate the application of BIM. In this regards, this research tried to address the interoperability problems with a machine learning-based method by automatically enhancing the semantic information of the BIM model. In this paper, we focused on the semantic enrichment of space objects by using textual data in the IFC. We implemented the NLP-based classification training model, which employs the word embedding techniques. As an early phase of research, this paper conducted training experiments with variable input properties, name and area of space, from space object in Korea school buildings. The accuracy of the classification model with a single feature (name) is measured at 85.9%, which is higher than multi feature (name and area). The suggested model can be expanded to more variable properties and target objects as a part of the semantic enrichment system.

Keywords –

Building information modeling (BIM), Spatial information enrichment, Natural language processing (NLP), Industry foundation classes (IFC)

1 Introduction

The adoption of building information modeling (BIM) contribute to enhancing the efficiency of architecture, engineering, and construction (AEC) industry by generating and managing building information with a computational model. BIM enables the automation of variable processes such as code compliance checking, design analyses [1]. Industry Foundation Class (IFC) standard supports the data interoperability of BIM, which aims to maximize the advantage of BIM.

However, several problems still remain in terms of practical applications of BIM. Some basic analysis processes can be driven by quantitative data from IFC, but the practical analysis requires more complex and domain-specific data that are not defined in IFC scheme. The semantic information required for specific analysis tasks can be easily obtained by a human who has domain-specific knowledge. Domain experts reason the semantic information about building objects using various types of data, such as text, images or quantity values. In order to intelligently automate these reasoning process by a computer, there have been some researchers to implement the automated information inference system. The research area, called semantic enrichment, addressed the interoperability of semantic BIM data by employing rule-based or machine learning methodologies [2, 5-8]. This research is also a part of them, which tried to implement the semantic enrichment system focusing on a space object instance. In this paper, we propose the utilization of textual values to intelligently inference the semantic information of space objects, as domain experts.

As machine learning and deep learning technology have been dramatically developed, natural language processing (NLP), which is sub-domain of artificial intelligence, also has made an amount of progress [3]. Machine learning-based NLP enables the computer to learn the semantic meaning of natural languages by itself. Word embedding technique helps to represent the semantic meaning of words in numerical vectors, which make a breakthrough in variable NLP tasks. In this research, the NLP techniques including word embedding are employed to implement the classification model. The approach to NLP-based feature generation and spatial category classification model are depicted in this paper. The results of training experiments with spaces in Korean school buildings are also described for validations. The overview of this research and scope of this paper are illustrated in Figure 1.

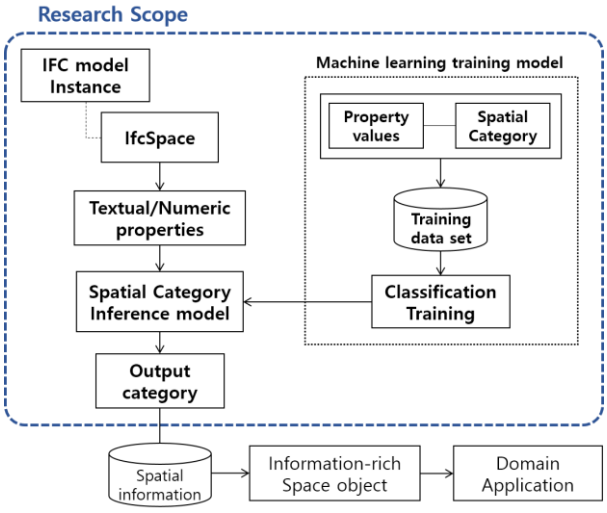

Figure 1. An overview of the suggested spatial information enrichment system and scope of this paper

2 Background

2.1 Semantic information of building objects for domain-specific application

As a standardized data specification, IFC is meant to support data interoperability between the different BIM platforms, however, there have been some studies reported practical problems of IFC [3]. Domain-specific applications require higher levels of semantic information and the required information is different for each domain task. Model View Definitions (MVDs) was suggested to define a subset of IFC required for domain-specific applications. However, MVD approach has a limitation on practical applications since it has to be defined respectively for every application.

Recent studies in code compliance checking, building analysis, and BIM model query tried to develop a system to supplement the semantic information. Solihin et al. suggested FORNAX platform for code compliance checking system, which supplements the semantic attributes of building objects with pre-defined FORNAX objects [5]. Lee et al. developed space database and mapping algorithms that provides the standard information of space object instances [6]. Belsky et al. proposed a Semantic Enrichment Engine for BIM (SEEBIM) which is a rule-based inference system for classifying components of bridge models [7].

Tanya and Rafael suggested machine learning-based inferencing method for semantic enrichment of space objects. Using the single features and pairwise features of space objects, the machine learning algorithm is trained to classify the type of room [2]. Koo et al. utilized support vector machines, which is one of the machine learning algorithms, to classify the building objects and check the semantic integrity of them [8]. These studies tried to switch the rule-based inference method into machine learning-based method, and showed the potential of the approaches.

2.2 Spatial category classification

Space is one of the critical elements in the computer-based information system for the concept design, construction process, and facility management process [5, 8]. As a functional element where human activities are performed, space is used as a unit of design or building analysis.

Traditional 2D CAD system only represents space object with surrounding physical objects and labels. BIM application is based on the objects-based modeling, which represents space object dependent on their

Table 1. Classification of spatial categories and training labels for Korean school buildings [11, 12]

Space use category	Training label	Space use category	Training label
Classroom	1. General classroom	General Use Facilities	13. School cafeteria
Laboratory Facilities	2. Laboratory room		14. Corridor
Office Facilities	3. School office		15. Lobby
	4. Principal's office		16. Toilet
	5. Counseling room		17. Parking lots
	6. Teacher's lounge	Support Facilities	18. Security office
Study Facilities	7. Library		19. Electric room
	8. Computer lab		20. Water tank room
Special Use Facilities	9. Gymnasium		21. Machine room
	10. Practice room	Health Care Facilities	22. Nursing room
	11. Audiovisual room		
	12. Multi-purpose room		

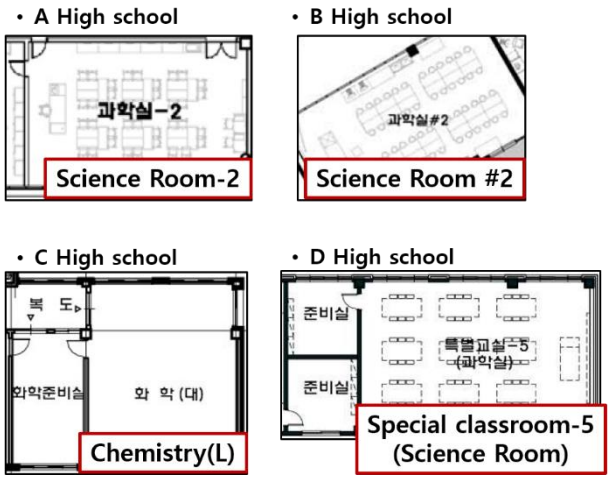

Figure 2. Examples of different naming convention for laboratory classroom

physical boundaries. This approach enables AEC industry stakeholders to manage and utilize required information of space objects.

Although, BIM provides rich-information about space objects, there are limitations on the interoperability of space objects in the domain application. Space instance can be classified differently, according to the applications. In some analysis, several spaces can be defined in a single zone, which is an aggregation of spaces. These characteristics hinder the utilization of space objects and their associated information. As an early phase of research to address the classification problems, we propose an approach to automated spatial category classification for space instance in the BIM model using machine learning algorithms. The motivation of this research is based on the research of Lee et al [6], which used space name data for classifying spatial categories.

3 Spatial categories classification of space objects for school buildings in Korea

The environmental comfort of educational facilities, which primarily depends on the building design, can affect the work or study productivity [10]. In this regards, many countries make efforts to establish more comfortable and effective facilities for both students and teachers. The Korean government also tries to improve the education environment with advanced facilities. The standard area for each student and classroom is specified in design regulations. In addition, the Korean ministry of education tried to introduce subject classroom system, which requires classrooms for each subject [11]. The design plans of school buildings have to follow these guidelines. In terms of subject classroom system, classifying the spatial categories of spaces is critical for the design and assessment of school building.

The fundamental features for spatial category classification is a name and area. The problem of utilization of name property is that the computer cannot understand the textual semantics of the space names written by different naming conventions. Figure 2 shows a simple example of naming convention problems. The floor plans of each project have a different naming convention for the same subject classroom.

In this paper, we focused on the functional usage of space object among the spatial categories. The classification of usage is listed in Table 1. Omniclass Table 13: Spaces by Function [12] and Classification system in Subject Class Operation Guidelines in Korea [11] are considered for the space classification system. The concept of high-level categories is borrowed from Omniclass and detailed training labels referenced the classification in Korean guidelines.

4 Development of NLP-based spatial category classification model

4.1 Property Extraction from IfcSpace object

The classification model in this research is based on a supervised learning method, which requires input features and corresponding label of objects. The input feature is a critical factor for machine learning since the model is trained to capture the patterns of input features and use the patterns for classification. The property values of IfcSpace objects are extracted for generating input feature of the usage classification model. The extracted property values of each IfcSpace instance object are expressed in a single vector as an input of the classification model.

Table 2. A list of IfcSpace properties

Attribute /Pset	Property name	IFC entity	Data type
Attribute	LongName	IfcLabel	String
	Name	IfcLabel	String
	GlobalId	IfcGolbally UniqueID	String
	Description	IfcText	String
	...	...	...
Property set	Department	IfcText	String
	Occupant	IfcText	String
	Area	IfcArea Measure	Numeric
	Unbounded Height	IfcLength Measure	Numeric
	...	...	...

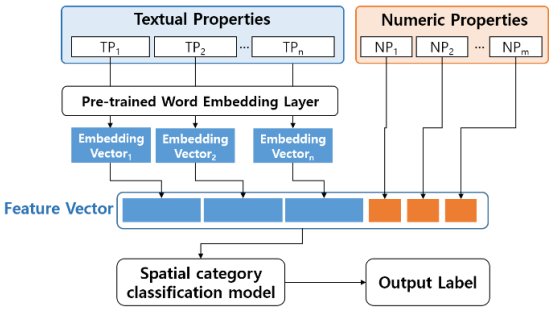

Figure 3. Feature extraction from space object properties

Table 2 shows a snippet list of IfcSpace properties. The IfcSpace object has attributes and property set which express the information of a space. Each property has a specific data type suitable for expressing the information. The feature generation is proceeded with converting property values in a suitable form. Numeric or Boolean data can be simply converted into input features. In order to translate textual data in a numeric form, word embedding techniques are applied. The details of utilization of word embedding are described in the next section. In this paper, we use the values of space name and area among the listed properties for classification.

4.2 Feature generation with NLP techniques

Neural network-based word embedding techniques become a prominent method for representing text data into the numeric form. In order to help the computer to understand and process the natural language, there have been variable methods to encode text in numeric format. The count-based representation approach has a limitation for representing the semantics of each word. Distributed representation makes a breakthrough by predicting the possibility of co-occurrence of words or phrases and encoding their features into vector format [13]. The concept of distributed representation is recently implemented with neural network-based models, which can handle the amount of text data. In this research, we obtained word vectors with word embedding model named fastText developed by Facebook's AI Research (FAIR) lab. The fastText model uses subword information for obtaining word vectors, which enables a model to represent the words that did not appear in the training data set [14].

We train the 100-dimensional word embedding model with 356,010 sentences from Korean Act sentences collected with an Open API data of National Law Information Center (law.go.kr). The 100-dimensional word vectors can be used for classification alone or concatenated with other numeric properties (Figure 3). Other numeric properties are obtained from the property value of given space object, with IFC parsing tool such as IfcOpenShell or it can be simply derived from the room schedule exported from BIM authoring tool. An approach to the explicit representation of relational properties such as connection, inclusion, adjacency is also another important issue for utilizing the information to machine learning or deep learning training.

4.3 Classification training and inference model

The classification model is implemented with a feed-forward network composed of three hidden layers. Each hidden layer contains 50 nodes respectively and the size of the output layer is 22 which is same as the number of output labels. ReLU (Rectified Linear Unit) function is used for activation function of each layer. The cross-entropy loss function and the stochastic gradient descent algorithm are employed to calculate the loss of training and optimize the model. The neural network model is implemented with the PyTorch framework [15]. The inference model makes an inference for new input data with a classifier of the trained model.

5 Training experiments and results

5.1 Training experiments

The training data set is collected from the best practice examples of educational facilities in Korea, provided by Education Facilities Research & Management Center (EDUMAC). EDUMAC provides design proposals of school buildings which include floor plans and space programs. In this research, 7 buildings are selected by the author, and 598 space data are collected for training. Training data is randomly separated in 80:20 ratio for training data and validation data, thus 478 spaces are used for training and 120 spaces are used for measuring the accuracy of trained data.

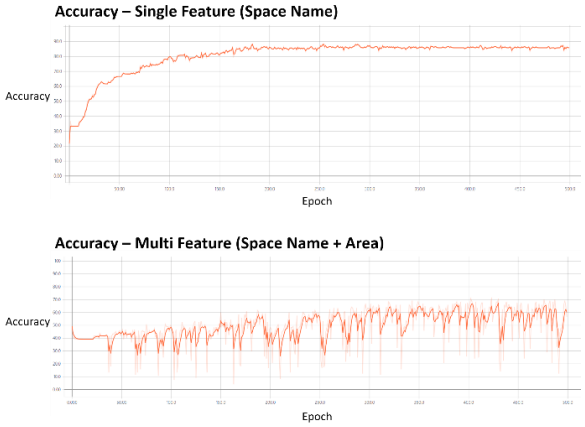

Figure 4. Training accuracy of each training

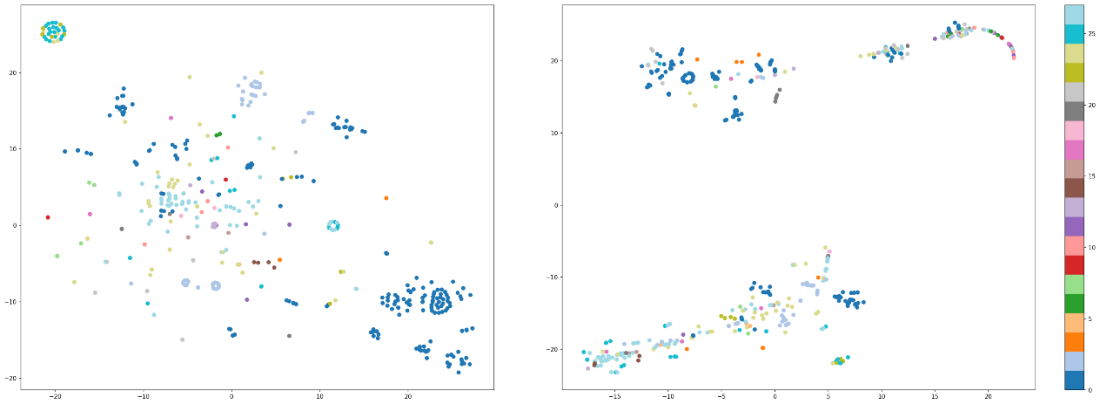

Figure 5. t-SNE Visualization of input feature vectors; Left: 100-dimensional vector (word embedding vector), Right: 101-dimensional vector (concatenated vector of word embedding and space area)

Training experiments are conducted with two scenarios. The first training is proceeded with a single feature, in this case, word embedding vectors of space name. In the second training experiment, the multi feature, which is a concatenated vector of word embedding vector and the value of the space area, are inputted to a training model. The training model and the number of training epoch are used for training, to compare the results.

5.2 Training results

The training results of each training model with 500 iterations are illustrated in Figure 4. Figure 4 illustrates the plot of training accuracy of single feature training model in the above graph and multi feature training model below. As shown in the above graph, there are no advances in training after 350 times of iteration for the single feature training result. The final training accuracy is measured by 85.9%. The training accuracy of second training experiment with multi feature is measured lower than the former one, measured to 60%. This result shows one single feature - space area in this case - can affect the training outcome by changing the input values and even produce worse results.

Figure 5 is a t-SNE (t-distributed Stochastic Neighbor Embedding) visualization [16] of input feature vectors. t-SNE visualization is a visualization tool for high dimensional data, which capture the similarity between the points and tries to express them in a lower dimension. By visualizing data in 2D or 3D space, the user can easily figure out the distributed pattern of data. In Figure 4, Left one is a 2D scatter plot of the 100-dimensional vectors with a single feature and the right one is the plot of 101-dimensional vectors with the multi feature. The color of dots represents the usage label of each observation. The clusters of data are more easily identified in left scatter plot than the right one. In other words, the computer can easily figure out the patterns and classify input observations of single feature vectors than multi feature.

This result can be attributed to the design plan of school building based on the unified spatial module. Most of the spaces in school buildings have a similar area and geometric feature regardless of functional features. They are partitioned to a certain size which is mostly based on the size of the general classroom. Whether the functional usage of space is laboratory classroom or teacher office, most of them have very similar, even almost the same, geometric features to the general classroom. In the training experiment from this research, the area value acts as a noise to lower classification accuracy since they cannot show the pattern to distinguish the usage. In the future work, other geometric properties and relational properties of space object can be appended to the input features for validation of the proposed approach.

6 Conclusion

This research proposes a spatial category classification method utilizing name and area of space objects. The spatial category classification is based on a neural network model, which use the word embedding techniques for processing textual data. The results of training experiments show that deep learning model can classify the space object by their name. However, deciding which properties to be utilized as training features is critical for training output.

The contribution of this research is to expand the utilization of textual data that is one of the important values for inferencing the semantic information of space object. In the proposed classification model, machine learning-based NLP techniques are employed, which can help the computer to understand the semantic meaning of natural languages. The machine learning-based model

can deal with the lexical problems, such as acronyms, omissions, which are difficult to handle in the rule-based algorithm.

This research could be expanded in terms of input features and target objects. Input features can be expanded with other textual data or numeric data. If there is more information about occupants or department of space, these textual properties could be converted into the feature vector of training models. Additionally, the other geometric properties and relational properties also can be applied to training. Utilization of variable properties are expected to be more independent to the requirements of input properties; Even if some information is missing in the BIM model, the trained model may infer the semantic information about the space objects. The target of the classification also can be expanded to variable objects including building itself.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2019R1A2C1007920).

References

[1] Eastman, Chuck, et al. BIM handbook: A guide to building information modeling for owners, managers, designers, engineers and contractors. John Wiley & Sons, 2011.

[2] Bloch, T. and Sacks, R. Comparing machine learning and rule-based inferencing for semantic enrichment of BIM models. *Automation in Construction*, 256–272, 2018.

[3] Young, T., Devamanyu, H., Poria, S., and Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55–75, 2018.

[4] Pazlar T. and Turk Z. Interoperability in practice: geometric data exchance using the IFC standard, *Journal of Information Technology in Construction* (*ITcon*), 13, Special issue Case studies of BIM use: 362-380, 2008.

[5] Solihin, W., Shaikh, N., Rong, X. and Lam, K. Beyond interoperability of building model a case for code compliance checking, 2004.

[6] Lee, J. K., Lee, J., Jeong, Y. S., Sheward, H., Sanguinetti, P., Abdelmohsen, S. and Eastman, C. M. Development of space database for automated building design review systems. *Automation in Construction*, 24: 203–212, 2012.

[7] Rafael, S., Ling, M., Raz, Y., Andre, B., Simon, D. and Uri, K. Semantic Enrichment for Building Information Modeling: Procedure for Compiling Inference Rules and Operators for Complex Geometry. *Journal of Computing in Civil Engineering*, 31(6): 4017062, 2017.

[8] Koo, B., La, S., Cho, N.-W. and Yu, Y. Using support vector machines to classify building elements for checking the semantic integrity of building information models. *Automation in Construction*, 98:183–194, 2019.

[9] Ekholm, A. and Fridqvist, S. A concept of space for building classification, product modelling, and design. *Automation in Construction*, 9(3): 315–328, 2000.

[10] da Graça, V. A. C., Kowaltowski, D. C. C. K. and Petreche, J. R. D. An evaluation method for school building design at the preliminary phase with optimisation of aspects of environmental comfort for the school system of the State São Paulo in Brazil. *Building and Environment*, 42(2): 984–999. 2007.

[11] Korean Educational Development Institute, Manual for subject classroom system 2018 – middle school, Research articles CRM 2018-146, 2018.

[12] OmniClass, OmniClass Construction Classification System, http://www.omniclass.org/, Accessed date: 30, January, 2019.

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26: 3111-3119, 2013.

[14] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

[15] PyTorch, Depp learning platform (2018) (https://pytorch.org/ accessed January 31, 2019)

[16] Maaten, Laurens van der and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research* 9: 2579-2605, 2008.