Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

Automated Design Information Extraction from Construction Specifications to Support Wood Construction Cost Estimation

Temitope AKANBI[1] and Jiansong ZHANG, Ph.D., A.M.ASCE[2]

[1] Graduate Student, Automation and Intelligent Construction Lab (AutoIC), School of Construction Management Technology, Purdue University, West Lafayette, IN. 47907; email: takanbi@purdue.edu

[2] Assistant Professor, Automation and Intelligent Construction Lab (AutoIC), School of Construction Management Technology, Purdue University, West Lafayette, IN. 47907; email: zhan3062@purdue.edu

ABSTRACT

In achieving full automation of construction cost estimation, the complete processes involved in computing cost estimates must be automated. The typical processes involved in achieving cost estimates are: (1) classification and matching of model elements to their various categories; (2) taking off quantities from design documents or Building Information Models; (3) retrieving unit cost from a cost database; and (4) applying the unit costs and quantities in computing the cost estimate. Although, the level of automation in quantity takeoff has been relatively high, most commercial software programs still require manual inputs from estimators to: (1) match materials of building elements to work items; and/or (2) fulfill essential information requirements that may be missing from design models for accurate cost estimate computations. These missing information are usually obtained from the construction specifications in supplement to the design models. Automating the process of design information extraction from construction specifications can help reduce: (1) the time and cost of the estimation, (2) the manual inputs required in cost estimation computations, and (3) human errors in cost estimates. This paper explores the use of natural language processing techniques to help process construction specifications and the authors propose a new algorithmic method for extracting the needed design information from construction specifications to support wood construction cost estimation. A case study was conducted on a wood construction project to evaluate the authors' proposed method. The results showed that the proposed method successfully searched for and found design details from construction specifications to fulfil essential information requirements for detailed wood construction cost estimation, with a 94.9% precision and a 97.4% recall.

Keywords: Cost Estimation, Automation, BIM Interoperability, Wood Construction, Natural Language Processing.

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

INTRODUCTION

Building Information Modeling (BIM) offers new methods and techniques to improve the construction process and performance of buildings and infrastructures (Shou et al. 2015). However, while information required to plan, construct, operate and maintain construction projects could be automatically obtained from BIM in theory; practically, the information obtained from BIM can be limited (Lee et al. 2014). This is due in part to the different level of details (LOD) that can be created in BIM. An LOD specification provides a reference that defines the BIM information level of details (Choi et al. 2015). There are six fundamental LOD definitions: LOD 100, LOD 200, LOD 300, LOD 350, LOD 400 and LOD 500 (BIMForum 2017). According to the American Institute of Architects (AIA), models at LOD 400 include assembly details such as size, shape, location, quantity, orientation, fabrication, and installation information (BIMForum 2017).

In the construction cost estimation domain, some of the information required for cost estimation are obtained from a model at LOD 400 or greater. Because of the unavailability of all models at LOD 400 or greater for cost estimation purposes, estimators constantly have to manually extract design information from design specifications, outline specifications, and construction specifications, which are provided at the schematic design stage, design development stage, and construction stage, respectively (Charette and Marshall 1999). According to Nassar (2012), some designs are still represented in 2D views. As such, at times automated information extraction from complete 3D models cannot be easily achieved. These manual processes of extracting design information from specifications requires a thorough and in-depth construction/engineering knowledge (Staub-French et al. 2003; Ma et al. 2016). This contributes to a widely accepted knowledge that manual cost estimation is a tedious, time-consuming, and cumbersome task that usually involves human errors (Akanbi and Zhang 2017).

BACKGROUND

In preparing cost estimates, estimators utilize different cost breakdown structures (CBS) or work breakdown structures (WBS), which are based on the different construction classification systems available. Construction classification systems are procedures to organize and catalogue the built environment (Autodesk 2017). In the United States, the General Services Administration (GSA) mandates that cost estimates for projects should be reported using the UniFormat II or MasterFormat classification systems (The General Service Administration 2017).

Construction Specifications

According to Charette and Marshall (1999), design specifications contains preliminary project description and are provided at the schematic design stage. Outline specifications contains design project description and are provided at the design development stage. Construction specifications contains construction related

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

project description and are provided at the construction stage. In accordance with "CSI/180 FF practice guide," the Construction Specifications Institute (CSI) recommends the preparation of preliminary project description using the UniFormat II elements classification system and the use of the MasterFormat classification system for the preparation of outline and construction specifications, respectively (Charette and Marshall 1999). This paper focuses on retrieving design information from the MasterFormat classification system for preparing construction cost estimations.

MasterFormat Construction Specifications

MasterFormat provides a master list of numbers and titles for classifying construction work results, requirements, products and services (Autodesk 2017). MasterFormat is an accepted industry standard for building design and construction in North America since the 1960s (Afsari and Eastman 2016). MasterFormat has 50 divisions (divisions 0 through 49) and is generally used in the construction industry in the United States for: (1) writing specifications; (2) organizing bidding documents, contract documents, and product information; and (3) facilitating the communication among contractors, fabricators, specifiers, and suppliers (CSI 2016). In MasterFormat, construction specifications are categorized using a four-level hierarchical structure (Figure 1). The requirements for a project are divided into divisions, which are further divided into sections to provide details about the requirements of a specific product, material or activity.

DIVISION 6 - WOOD & PLASTICS	LEVEL 1
Section 061000 Rough Carpentry	LEVEL 2
PART 1 - GENERAL	LEVEL 3
1.1 SUMMARY	LEVEL 4

Figure 1. MasterFormat structure of construction specifications

Natural Language Processing (NLP)

NLP techniques simplifies the processing (understanding, analyzing, manipulating, generating) of natural languages (text or speech) using computers in a manner similar to human-processing (Cherpas 1992). NLP can be utilized to process text, e.g., to detect, locate, and extract specific pieces of information from the text. Grishman (2010) defines information extraction as the process of analyzing, classifying and obtaining instances or targeted information from a textual document. This can be accomplished by identifying patterns in the data and developing extraction rules based on the patterns for extracting the important information, as shown by Zhang and El-Gohary (2016; 2017). According to Kaiser and Miksch (2005), there are two main ways of establishing an information extraction system: (1) the learning technique approach; and (2) the knowledge engineering approach. The learning technique approach utilizes a reverse engineering method to analyze data in a particular domain and define relevant text patterns that would be used in the extraction

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

of relevant information. The patterns are then used to develop rules which can be applied on unseen text data in the same domain. The knowledge engineering approach utilizes an iterative process in developing rules to extract the relevant information using an available corpus (a library of texts) of "relevant" texts. The authors' proposed extraction system was developed using the learning technique approach to create rules from a set of design specifications/construction specifications which served as the training data.

PREVIOUS RESEARCH EFFORTS

In an effort to automate the manual processes involved in construction cost estimation, Staub-French et al. (2003) proposed a feature-driven activity and resource classification system for predicting construction costs by extracting and matching the activity specifications of a component. Lee et al. (2014) proposed a method utilizing a semantic reasoning mechanism to automate the inference of work conditions by extracting the work conditions and work items from the design information. Ma et al. (2016) developed an ontology-based approach for formalizing cost specifications in China to support an improved implementation in computer programs. These efforts improved the processes involved in construction cost estimation. However, these efforts are still limited in terms of achieving full automation of construction cost estimation. In contrast, the method proposed in this research aims to: (1) automatically extract all the cost parameters required for computing the cost estimates from design documents using a major classification system in the construction industry; and (2) automatically save the extracted parameters in a database that can be further utilized for identifying, matching and retrieving the unit cost of the material from the database.

PROPOSED METHOD

The authors proposed an automated design information extraction (IE) algorithm development method with the following four steps (Figure 2). Step (1): analysis – analyze the text of design and construction specifications for wood structures. This step includes text classification and preprocessing. Step (2): identification – detect the target pieces of design information to be extracted from the text. Step (3): IE algorithm development – after detecting the target pieces of design information to be extracted, the algorithms to extract the needed information from each section are developed. Several natural language processing techniques are utilized to support the pattern matching and extraction, including tokenization, sentence splitting, and morphological analysis. Tokenization breaks a text into tokens. Sentence splitting splits the text into individual sentences. Morphological analysis is necessary because of variations of word forms that might occur in the text. Morphological analysis is used to match these different forms of the same root word (i.e., semantic root), such as noun form, verb form, etc. Step (4): evaluation – evaluate the performance of the developed IE algorithms, the extraction results of the IE algorithms are compared to the results manually generated by estimators. Precision and recall are used to measure the performance of the IE algorithms. Precision is

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

defined as the number of correctly extracted design items divided by the total number of design items extracted. Recall is defined as the number of correctly extracted design items divided by the number of design items to be extracted in the specification.

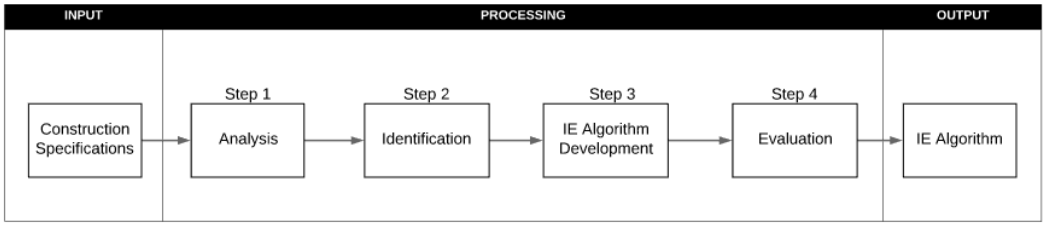


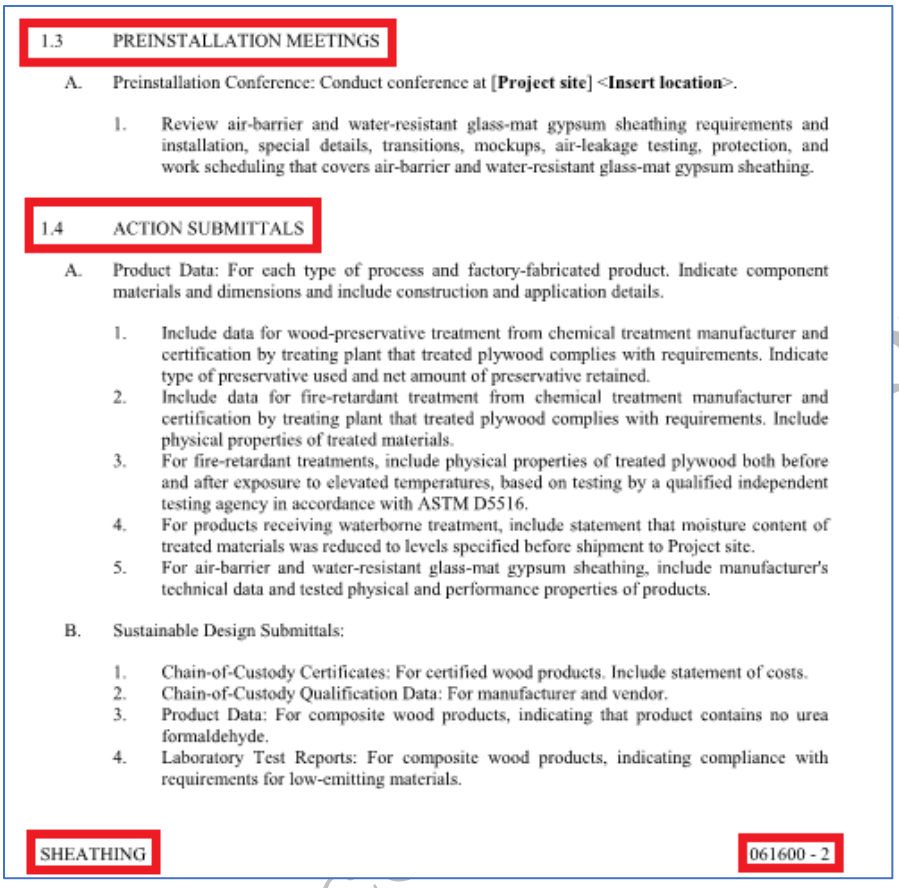
Figure 2. Proposed automated design information extraction (IE) algorithm development method

EXPERIMENTAL RESULTS AND ANALYSIS

The proposed automated design information extraction algorithm development method was tested in an experiment to develop IE algorithms for retrieving the cost design parameters for a residential wood structure. The experimental details are described as follows.

Step 1 - Analysis: In this step, the construction specification texts were classified and preprocessed. A construction specification document contains much information, some of which do not pertain to cost estimation directly. As an example, Sub-sections 1.3 and 1.4 of "SECTION 061600 – SHEATHING" from "DIVISION 06 – WOOD & PLASTIC" of the AIA construction specification typically contain information related to meetings and submittals (Figure 3). As shown in Figure 3, Section 1.3 provides information related to pre-installation meetings while Section 1.4 provides information related to mandatory submittals (product data and sustainable design submittals). A text classification was performed to decipher the relevant text for extraction.

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069



1.3 PREINSTALLATION MEETINGS

A. Preinstallation Conference: Conduct conference at [**Project site**] <**Insert location**>.

1. Review air-barrier and water-resistant glass-mat gypsum sheathing requirements and installation, special details, transitions, mockups, air-leakage testing, protection, and work scheduling that covers air-barrier and water-resistant glass-mat gypsum sheathing.

1.4 ACTION SUBMITTALS

A. Product Data: For each type of process and factory-fabricated product. Indicate component materials and dimensions and include construction and application details.

1. Include data for wood-preservative treatment from chemical treatment manufacturer and certification by treating plant that treated plywood complies with requirements. Indicate type of preservative used and net amount of preservative retained.
2. Include data for fire-retardant treatment from chemical treatment manufacturer and certification by treating plant that treated plywood complies with requirements. Include physical properties of treated materials.
3. For fire-retardant treatments, include physical properties of treated plywood both before and after exposure to elevated temperatures, based on testing by a qualified independent testing agency in accordance with ASTM D5516.
4. For products receiving waterborne treatment, include statement that moisture content of treated materials was reduced to levels specified before shipment to Project site.
5. For air-barrier and water-resistant glass-mat gypsum sheathing, include manufacturer's technical data and tested physical and performance properties of products.

B. Sustainable Design Submittals:

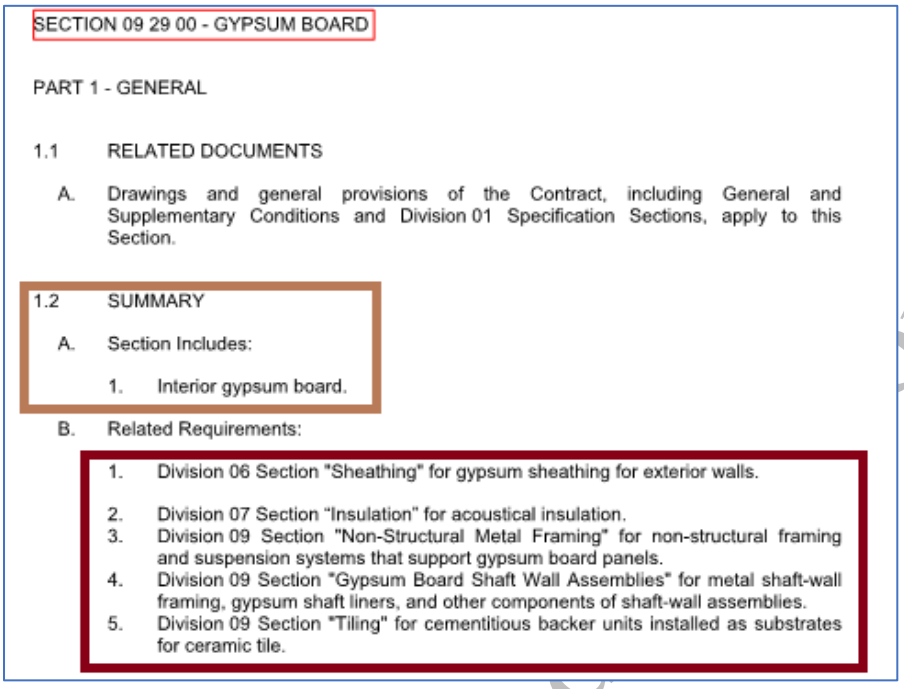
1. Chain-of-Custody Certificates: For certified wood products. Include statement of costs.
2. Chain-of-Custody Qualification Data: For manufacturer and vendor.
3. Product Data: For composite wood products, indicating that product contains no urea formaldehyde.
4. Laboratory Test Reports: For composite wood products, indicating compliance with requirements for low-emitting materials.

SHEATHING 061600 - 2

Figure 3. Section 061600 – sheathing (Part 1, Sections 1.3 & Section 1.4) of the AIA construction specification

Step 2 - Identification: In this step, the analyzed texts from Step 1 were further processed for identifying the target information. Construction specifications are semi-structured, that is, in spite of the metadata they main contain, the documents still lack a detailed organizational structure, and the level of detail may vary significantly based on a designer's preferences or an organization's standard of practice. Even after filtering irrelevant texts, the relevant texts are still mixed and diverse. Two or more divisions, sections, or subsections could be interlinked. Furthermore, a sentence in a construction specification may contain information required for two or more cost parameters (Figure 4). As an example, as shown in Figure 4, "SECTION 092900 – GYPSUM BOARD," which contains information pertaining to gypsum board, is related to "SECTION 061600 – SHEATHING," "SECTION 072100 – THERMAL INSULATION," "SECTION 092216 – NON-STRUCTURAL METAL FRAMING," and "SECTION 093013 – CERAMIC TILING." Furthermore, a gypsum board finish level in a construction specification can include two or more finish levels of gypsum board locations (e.g., ceiling plenum areas and exposed wall areas) within a sentence (Figure 5).

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069



SECTION 09 29 00 - GYPSUM BOARD

PART 1 - GENERAL

1.1 RELATED DOCUMENTS

A. Drawings and general provisions of the Contract, including General and Supplementary Conditions and Division 01 Specification Sections, apply to this Section.

1.2 SUMMARY

A. Section Includes:

1. Interior gypsum board.

B. Related Requirements:

1. Division 06 Section "Sheathing" for gypsum sheathing for exterior walls.
2. Division 07 Section "Insulation" for acoustical insulation.
3. Division 09 Section "Non-Structural Metal Framing" for non-structural framing and suspension systems that support gypsum board panels.
4. Division 09 Section "Gypsum Board Shaft Wall Assemblies" for metal shaft-wall framing, gypsum shaft liners, and other components of shaft-wall assemblies.
5. Division 09 Section "Tiling" for cementitious backer units installed as substrates for ceramic tile.

Figure 4. Section 092900 – gypsum board (Part 1, Section 1.2, Paragraphs A & B)

Step 3 – IE Algorithm Development: In this step, the IE algorithms were developed to extract the target texts from the construction specification document. After the cost-related design information to be extracted were identified, the algorithms to extract these information were manually developed. Python language was utilized for the implementation of the extraction algorithms. Tokenization, morphological analysis and sentence splitting were utilized to support the pattern-based matching of IE rules. Tokenization breaks a text into tokens. As an example, after tokenizing Paragraph G.1 from Figure 5, the text segment "Level 1: Ceiling plenum areas, concealed areas, and where indicated." became fourteen tokens – 'Level' '1' ':' 'Ceiling' 'plenum' 'areas' ',' 'concealed' 'areas' ',' 'and' 'where' 'indicated' '.' Through morphological analysis, "tiles" would match "tile" in a string search. The IE algorithm developed for "SECTION 092900 – GYPSUM BOARD" is shown in Figure 6. The algorithm includes 10 processes. Processes 1 initiates a string search for divisions based on division label from the provided work breakdown structure input. Processes X, Y & Z are natural language techniques used to enhance the efficiency and coverage of the searches. Decision 1 checks if there is a division match found in the specifications. If there is a division match, the algorithm proceeds to Process 2. Processes 2, 3 & 4 are similar in that they initiate string searches while Decisions 2 & 3 checks if there are matches found in the specification. If Decisions 2 & 3 found matches in the specifications, Processes 5 – 10 extract the design information needed for cost estimation computations and write them into the following material cost parameters: (1) Type; (2) Thickness; (3) Name; (4) Use; (5)

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

Finish_Level; (6) Location. These cost parameters would be used to automatically extract the unit cost of the material from a database (Figure 7).

G. Gypsum Board Finish Levels: Finish panels to levels indicated below and according to ASTM C 840:

1. Level 1: Ceiling plenum areas, concealed areas, and where indicated.
2. Level 2: Panels that are substrate for tile.

Figure 5. Section 092900 – gypsum board (Part 3, Section 3.1, Paragraph G)

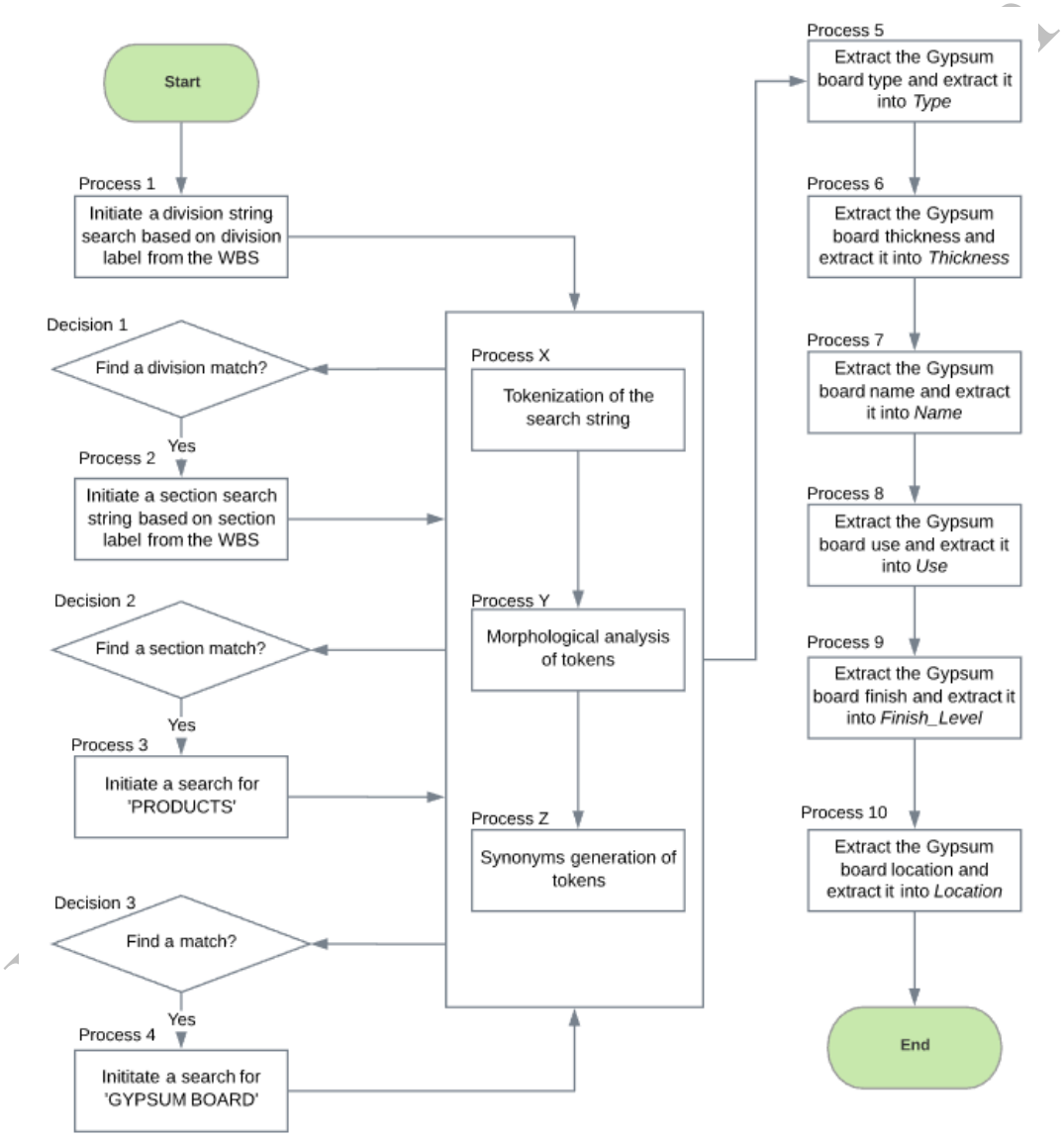


Figure 6. Flow chart of the developed IE algorithm for gypsum board

Step 4 Evaluation: In this step, the performance of the IE algorithms was tested on the construction specifications of a residential wood apartment project in Detroit,

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

Michigan. The project consisted of a two-story wood shell enclosure comprised of wall systems, floors, a roof system, a stair system, windows and doors. Two metrics, precision and recall, were used to evaluate the performance of the proposed IE algorithms. To create a gold standard, two professional estimators manually extracted cost parameters from the construction specifications. In total, 76 cost parameters were extracted from fifteen sections of the construction specifications. The resulted IE algorithm achieved 97.4% recall and 94.9% precision (Table 1). The results indicate that the proposed IE algorithm development method is promising in developing algorithms that can automatically extract cost parameters from construction specifications.

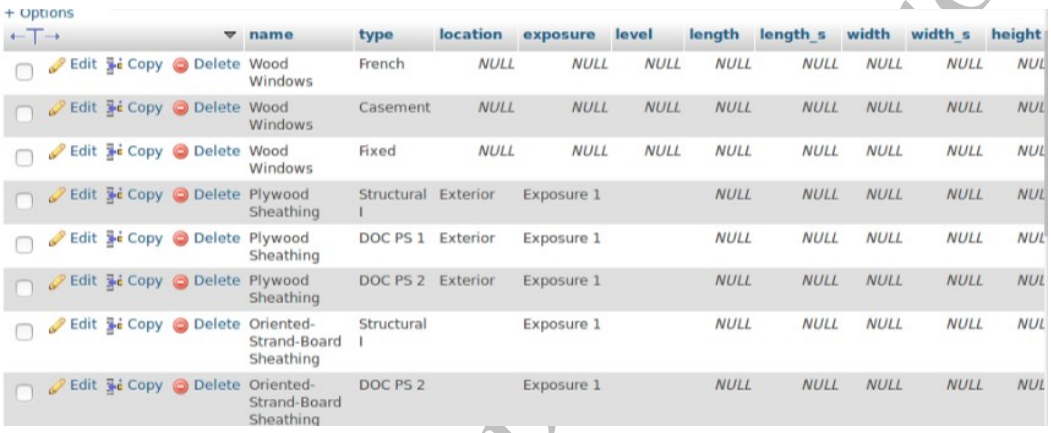


Figure 7. Material database results (partial)

Table 1. Experimental Results

	Division 06	Division 07	Division 08	Division 09	Total /Avg.
No. of parameters in Standard	12	11	11	42	76
No. of correctly extracted parameters	12	10	11	41	74
No. of totally extracted parameters	13	11	11	43	78

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

Precision	92.3%	90.9%	100%	95.3%	94.9%
Recall	100%	90.9%	100%	97.6%	97.4%

CONCLUSIONS AND FUTURE WORK

In this paper, the authors proposed a new method to develop IE algorithms that can automatically extract cost parameters from construction specifications in MasterFormat. The proposed method produces algorithms that write the extracted cost parameters in a database that can be utilized to automatically retrieve the unit cost of materials. The proposed method was tested in extracting the cost parameters from the construction specifications of a wood building project. The resulted IE algorithms achieved 94.9% precision and 97.4% recall, in comparison to a manually developed gold standard by professional estimators. The experimental results indicate that the proposed method can be used to generate algorithms that can automatically extract the cost parameters from construction specifications in MasterFormat.

The scope of this research was limited to wood construction. In future work, the authors plan to test the proposed method in extracting cost parameters from other divisions of the MasterFormat (e.g., concrete, steel) and other construction classification systems such as Uniformat II.

ACKNOWLEDGMENTS

The authors would like to thank the National Science Foundation (NSF). This material is based on work supported by the NSF under Grant No. 1745374. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

Akanbi, T., and Zhang, J. (2017). "Automated wood construction cost estimation" *Proc., 2017 ASCE Intl. Workshop on Comput. in Civ. Eng.,* ASCE, Reston, VA., 141-148.

Afsari, K., and Eastman, C. (2016). "A Comparison of Construction Classification Systems Used for Classifying Building Product Models" *Proc., 2016 ASC Annual International Conference,* Denver, CO., 101-108.

Autodesk. (2017). "Classification Systems and Their Use in Autodesk Revit: Managing the "I" in BIM." https://www.biminteroperabilitytools.com/classificationmanager/downloads/Autodesk%20Whitepaper%20-%20Classification%20Systems.pdf> (April 12, 2019).

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

BIMForum. (2017). "Level of development specification guide: November 2017." < https://bimforum.org/wp-content/uploads/2017/11/LOD-Spec-2017-Guide_2017-11-06.pdf> (Apr. 10, 2019).

Charette, R., and Marshall, H. (1999). "Uniformat II Elemental Classification for Building Specifications, Cost Estimating, and Cost Analysis." National Institute of Standards and technology (NIST).

Cherpas, C. (1992). "Natural language processing, pragmatics, and verbal behavior." *Analysis of Verbal Behavior*, 10,135-147.

Choi, J., Kim, H., and Kim, I. (2015). "Open BIM-based quantity takeoff system for schematic estimation of building frame in early design stage." *Journal of Computational Design and Engineering,* 2(2015), 16-25.

CSI. (2012). MasterFormat. https://www.edmca.com/media/35207/masterformat-2016.pdf> (July 22, 2019).

Grishman, R. (2010). *Information Extraction. The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell, United Kingdom, 515–530.

Kaiser, K., and Miksch, S. (2005). "Information Extraction: A survey."

Kim, T., and Chi, S. (2019). "Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry." *J. Constr. Eng. Manage.,* 2019, 145(3): 04019004.

Lee, S., Kim, K., and Yu, J. (2014). "BIM and ontology-based approach for building cost estimation." *Automation in Construction,* 41(2014), 96-105.

Ma, Z., Liu, Z., and Wei, Z. (2016). "A generic feature-driven activity-based cost estimation process." *Advanced Engineering Informatics,* 17(2003), 96-105.

Nassar, K. (2012). "Assessing building information modeling estimating techniques using data from the classroom." *Journal of Professional Issues in Engineering Education & Practice,* 138(2012), 171-180.

Shou, W., Wang, J., and Wang, X. (2015). "Formalized Representation of Specifications for Construction Cost Estimation by Using Ontology." *Computer-Aided Civil and Infrastructure Engineering,* 31(2016), 4-17.

Staub-French, S., Fischer, M., Kunz, J., and Paulson, B. (2003). "A generic feature-driven activity-based cost estimation process." *Advanced Engineering Informatics,* 17(2003), 96-105.

The General Services Administration (GSA). (2017). "PBS-P100: Facilities Standards for the Public Buildings Service." <https://www.gsa.gov/cdnstatic/2017_Facilities_Standards_%28P100%29%C2%A0.pdf> (April 14, 2019).

Zhang, J., and El-Gohary, N. (2016). "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking." *J. Comput. Civ. Eng.,* 2016, 30(2): 04015014

Final Published Version Should be found in the ASCE Database here:
Akanbi, T., and Zhang, J. (2020). "Automated design information extraction from construction specifications to support wood construction cost estimation." Proc., ASCE Construction Research Congress, ASCE, Reston, VA, 658-666. https://ascelibrary.org/doi/10.1061/9780784482889.069

Zhang, J., and El-Gohary, N. (2017). "Integrating semantic NLP and logic reasoning into a unified system for fully-automated code checking." Automation in Construction, 73, 45-57.