buildings

MDPI

Article

Towards Automated Construction Quantity Take-Off: An Integrated Approach to Information Extraction from Work Descriptions

Shengxian Tang [1], Hexu Liu [1,*], Manea Almatared [1], Osama Abudayyeh [1], Zhen Lei [2] and Alvis Fong [3]

[1] Department of Civil and Construction Engineering, Western Michigan University, Kalamazoo, MI 49008, USA; shengxian.tang@wmich.edu (S.T.); maneamohammeds.almatared@wmich.edu (M.A.); osama.abudayyeh@wmich.edu (O.A.)
[2] Department of Civil Engineering, University of New Brunswick, Fredericton, NB E3B 5A3, Canada; zhen.lei@unb.ca
[3] Department of Computer Science, Western Michigan University, Kalamazoo, MI 49008, USA; alvis.fong@wmich.edu
* Correspondence: hexu.liu@wmich.edu

Abstract: Construction-oriented quantity take-off (QTO) refers to the process of determining the quantities for construction items or work packages in accordance with their descriptions. However, the current construction-oriented QTO practice relies on estimators' manual interpretation of work descriptions and manual processes to look up proper building objects for quantity calculation. Hence, this research aims to develop natural language processing (NLP) and rule-based algorithms to automate the information extraction (IE) from work descriptions for QTO in building construction. Specifically, several named entity recognition (NER) models, including Hidden Markov Model (HMM), Conditional Random Field (CRF), Bidirectional-Long Short-Term Memory (Bi-LSTM), and Bi-LSTM+CRF, were developed to identify construction activities, material, building component, product features, measurement unit, and additional information (e.g., work scope) from work descriptions. Cost items in the RSMeans database are used to evaluate the developed models in terms of F1 scores. HMM was found to achieve a 5% higher F1 score in the NER than the other three algorithms. Then, labeling rules and active learning strategies were applied along with the HMM model, which improved F1 score by 3% and reduced the labeling efforts by 26%. The results showed that the proposed IE method successfully interprets the desired information from the work description for QTO. This research contributed to the body of knowledge by the NLP-based information extraction model integrating HMM and formalized labeling rules that automatically process work descriptions and lay a foundation for automated QTO and cost estimation.

Keywords: NLP; quantity take-off; cost estimation; construction automation; work description

Citation: Tang, S.; Liu, H.; Almatared, M.; Abudayyeh, O.; Lei, Z.; Fong, A. Towards Automated Construction Quantity Take-Off: An Integrated Approach to Information Extraction from Work Descriptions. *Buildings* **2022**, *12*, 354. https://doi.org/10.3390/buildings12030354

Academic Editor: Cinzia Buratti

Received: 25 January 2022
Accepted: 11 March 2022
Published: 15 March 2022

1. Introduction

Construction cost estimation is one of the most fundamental construction management tasks, intending to determine the total construction cost of projects before construction commences. It provides the base for cost management and control during the construction stage. Construction cost estimation typically involves several procedures [1], such as (1) developing construction methods, (2) establishing work breakdown structure (WBS), (3) take-off quantities for construction work packages in WBS, (4) calculating direct cost based on quantities and unit price of each work packages, and (5) determining the total construction cost by adding overhead, profit, and contingencies. However, these steps demand substantial manual efforts and are challenging to be fully automated. The reason for this partially arises from the fact that construction cost estimation is a knowledge-intensive process and estimation knowledge is missing from current computer systems. For example,

knowledge and experience of cost estimators are required to interpret the construction specifications to establish the WBS. The quantity take-off (QTO) step also demands manual judgments and involvements of estimators in analyzing work descriptions of the WBS cost items to determine their quantities accordingly. Construction-oriented QTO herein is defined as determining the quantity amount of construction cost items or work packages. As each cost item is associated with a specific construction crew and production rate, their unit cost regarding labor, material, and equipment varies. As such, each cost item is defined with a clear work scope through its unique work description. The work descriptions provide the basis for estimators in construction-oriented QTO and cost estimation.

Work descriptions are the textual information describing the nature and scope of work packages and construction tasks to deliver construction projects. Typical information of work descriptions includes construction material, construction method, product features such as locations and sizes, and accessories required. Construction-oriented QTO has to be determined in accordance with work descriptions of cost items. Work descriptions provide essential information regarding product and construction methods and are usually expressed using a collection of construction vocabulary, which are semi-structured and separated by commas. For example, a work description for a cost item of walls is '*Wall framing, studs, 2″ × 4″, 8′ high wall, pneumatic*' [2]. This description informs cost estimators to take off the total length of studs hosted by walls that are made of 2″ × 4″ studs, with a height of 8′, functioning as structural walls, and framed using a pneumatic nailing gun. Accordingly, the quantity for this cost item should be derived for building elements with the proper product features, namely, (1) quantity unit: length in the linear foot, (2) building material: lumber stud, (3) building element: wall, (4) material size: 2″ × 4″ for studs, (5) building element feature: wall height of 8′, and (6) building element feature: structural usage. Such information could guide estimators to extract the quantities of related building elements from the design document.

The traditional QTO is a tedious manual process that is subject to human error [3]. For example, substantive manual efforts from estimators are required to interpret the work descriptions manually in the traditional QTO. Different estimators may end up with different quantities results, even though they use the same work descriptions for cost items. The knowledge-based automation in QTO has been proven to be capable of addressing such identified issues. It can eliminate the manual measurement process, resulting in enhanced efficiency. Therefore, the knowledge-based automated QTO is required to address current issues in manual QTO. There is a need to extract desired information from work descriptions for automated construction-oriented QTO and construction cost estimation.

Although plenty of research has been devoted to NLP-based text analysis and information extraction in the construction industry, most of the existing literature primarily focuses on information extraction from inspection reports and construction specifications, consisting of natural language sentences. In contrast, work descriptions are a collection of construction vocabulary separated by commas and are not expressed with natural language sentence structure. At times, the collection of words for work descriptions has different sequential orders, i.e., unstructured data. For example, the work descriptions of wood framing activities for walls and floors are "*Wall framing, studs, 2″ × 4″, 8′ high wall*" and "*2″ × 6″* rafters, roof framing, to 4 in 12 pitch", respectively. Such variation in work descriptions imposes challenges in automated IE. In addition, the association of the desired information, such as building element/material and size, is difficult to identify, as the size can be either material size or building element size. These challenges affect the performance of the existing IE model. An automated approach that extracts the desired information from work descriptions for the purpose of construction-oriented QTO is lacking.

To fill this gap, this research developed an integrated approach for automatically extracting required information from the work description of cost items. Theoretically, the proposed approach contributed to the body of knowledge by integrating HMM and formalized labeling rules for automatically processing work item descriptions. This approach achieves the integration of named entity recognition (NER) and IE rules, leading

to a better performance in terms of precision, i.e., the F1 score. It lays a foundation for automated QTO and cost estimation. The practical contributions of the presented research are two-fold, including (1) increased automation by reducing massive manual efforts in the QTO process and (2) enhanced accuracy of information extraction and interpretation through eliminating manual subjective interpretation of work descriptions, especially for junior estimators. With that, the extracted information could be used to query a given BIM model to automatically extract the desired quantity and achieve the mapping between cost items and the BIM model in the future. The NER-rule-based approach for automatic information extraction is developed in this research as the first step in automated cost estimation. It also lays a foundation for automated QTO and cost estimation and sheds light on artificial intelligence (AI) applications for smart construction.

The remainder of this paper is organized as follows. In Section 2, previous research regarding NLP application in construction is reviewed to clarify the research gap. Subsequently, the research methodology is illustrated in Section 3 in detail. Section 4 presents the case study, as well as their results. The final section concludes the paper, highlighting the research contribution.

2. Literature Review

NLP has been extensively studied to facilitate various tasks in the construction industry, such as compliance checking, document management, and social media-based data analytics for construction applications in the past two decades [4–10]. It could be applied to address IE from work descriptions for construction-oriented QTO. As such, this section provides a comprehensive review regarding rule-based IE, ML-based IE, text classification, and information retrieval from BIM models.

2.1. Rule-Based Information Extraction

IE is to extract desired information from unstructured text data. In general, IE can be performed in two different approaches, including (1) the rule-based method and (2) the ML-based approach. A rule-based IE is used to extract predefined information based on pattern matching. It is widely employed in the NLP research community and the construction industry. For example, Lee et al. [11] attempted to apply rule-based IE in contract management. An NLP-based extraction model was developed to identify poisonous clauses in international construction contracts. They developed semantic rules such as "if-then" logic to extract the predefined information in contracts. Additionally, a construction-oriented lexicon was used to facilitate semantic matching; for example, different words that may have the same meaning semantically are determined to be identical. Ontology is also often used for semantic modeling and is applied in IE because it allows formalizing the terms, interrelationships, and properties of domain terms. In this respect, Zhang and EI-Gohary [12] used ontology to incorporate semantic features and developed extraction patterns for NLP-based IE. Similarly, Xu and Cai [13] proposed an ontology and rule-based NLP approach to extract utility information and interpret textual regulations. Their approach can deal with complex spatial relations and reasoning for compliance checking.

2.2. ML-Based Information Extraction

Although rule-based IE can provide high performance such as precision, it, on the other hand, suffers from the fact that it demands substantial manual efforts in rule development. It is also challenging to apply the developed rules to other applications. With the advancement of artificial intelligence (AI), machine learning (ML) is gaining momentum and is increasingly used in many discipline-specific applications. A large amount of research has been devoted to applying ML to process unstructured text data, including bridge inspection reports [14,15], work descriptions [16–18], and construction specifications [19,20]. For example, Liu and EI-Gohary [14] invented an ontology-based semi-supervised CRF for extracting bridge deficiencies and maintenance actions from bridge inspection reports.

Their approach intends to reduce manual efforts in rule development of the rule-based approach and data labeling of ML-based IE. Similarly, Kim and Chi [21] used the rule-based method and conditional random field (CRF) to extract tacit knowledge such as hazard objects, hazard position, work process, and accident results from accident cases. The CRF model was trained using rule-based labeled data. Integrating rule-based methods and ML can reduce the manual efforts in rule development and data labeling. Subsequently, Liu and EI-Gohary [15] further proposed a novel semantic neural network ensemble-based method to identify semantic dependency relations of extracted information in bridge reports. Another effort is the NLP-based text analysis for assigning maintenance staff for building maintenance [22]. In their research, an ML-based classification model is trained to assign and prioritize work orders for building maintenance, achieving 77% and 88% accuracy for staff assignment and prioritization, respectively.

Moon et al. [19] reported an NER model to extract user-defined information from construction specifications, with a particular focus on road construction projects. The NER model is essentially a Bi-LSTM model, and their research is one of the first few attempts in successfully applying this model in the construction industry. As revealed by existing literature, the ML-based approach for IE demands significant manual efforts in labeling the training data. In this regard, Moon et al. [23] proposed to integrate active learning and a recurrent neural network (RNN) for bridge damage recognition. The results proved that the proposed model is capable of detecting bridge damage with reduced effort.

2.3. Text Classification

Alternatively, Martínez-Rojas et al. [16–18] explored six classification methods to assign work descriptions to a predefined structure of task groups. These six methods include the C4.5 decision tree, random forest, Naïve Bayes, neural networks, support vector machines, and k-nearest neighbors. Basic linguistic processing, such as cleaning and synonym replacement, was applied before applying these classification methods. The results revealed that random forest achieves the best performance in terms of precision. Their proposed approach can organize construction data such as bills of quantities and work descriptions in a structured manner so that construction data can be readily accessed during the decision-making process. In terms of contract management, Jallan et al. [24] applied NLP-based text mining to reveal patterns in litigation cases regarding construction defects, i.e., the similarity of keyword frequency and topic modeling. Le and Jeong [6] explored an NLP-based approach to identify semantic relations of transportation asset data terminology. Their classification of domain terms lay a foundation for integrating asset data for the transportation industry.

2.4. Information Retrieval from BIM

In fact, some scholars used NLP in text analysis to facility QTO and cost estimation in construction. For example, Akanbi et al. [25] applied the NLP technique to analyze BIM models such as material layer information and then used identified material information to search a material database for price information. Akanbi et al. [20] presented an automated IE approach for extracting design information from construction specifications for cost estimation. Their system primarily extracts design information and then uses the extracted data to search unit prices in a material database for direct cost estimation. These efforts applied NLP to search cost items in cost databases instead of BIM models. On the contrary, Lin et al. [26] applied the NLP technique to information retrieval from a given BIM model. NLP was applied to interrupt the intent of end-users through the concepts of "keyword" and "constraints" in the query statement, and then, the desired information can be extracted from the BIM model. Wu et al. [27] employed the natural language processing technique to understand the user's query intention, leading to increased precision and efficiency of information retrieval from BIM models. Their approach primarily retrieves the product information from a given BIM model and is not intended for QTO for construction work packages. Liu et al. [28] proposed a knowledge model-based framework to calculate

QTO-related information through the developed standard method of measurement rules. Despite of this, the current interpretation of work descriptions of cost items is manual, time-consuming, and error-prone. Yet, there is a lack of an automated approach to extract desired information from work descriptions for the QTO purpose.

3. Research Methodology

To address these limitations, this research explores various NER models and rule-based methods to extract the product and process information from work descriptions for QTO. It is worth noting that NER is an IE task, where desired entities are identified from unstructured text and assigned with predefined labels [19]. This research addresses IE as a sequence labeling problem through a novel NER-based framework. The proposed sequence labeling framework can classify all target entities related to a given source entity in the cost item's work descriptions into predefined labels, such as construction material, construction method, function, size, sub-component of building elements, etc. The proposed framework achieves the integration of ML-based IE, rule-based approach, and active learning for IE from work descriptions, which reduces the label efforts and rule development while improving the IE performance, such as precision.

Figure 1 illustrates the research methodology. It consists of seven steps, namely, (1) data collection and preprocessing, where the authors collected cost items and retrieved the corresponding description from the construction cost database (i.e., RSMeans Online); (2) identification of labels, which is intended to determine what sort of information should be extracted for the purpose of construction-oriented QTO; (3) data labeling and preparation, which prepare training data and testing data for ML-based sequence labeling; (4) development and performance evaluation of NER algorithms, where four NER algorithms are developed and evaluated quantitatively in terms of F1 score; (5) formalization of sequence labeling rules, which function as prior knowledge to improve the performance of ML-based NER models; and (6) application of active learning to train the selected NER model (i.e., HMM), which reduces manual efforts spent on data labeling. Each of these steps is described in detail in the following sub-sections.

3.1. Data Collection and Preprocessing

Construction work packages or cost items are the basic units used to estimate the direct cost for construction projects. Each work package in cost databases has its description. For example, RSMeans online consists of thousands of cost items for construction estimation, and each item is associated with a unique description. Items in other cost databases, including the in-house database of contractors, also offer work descriptions. RSMeans online was selected in this study as it is the most widely used commercial cost database in North America [29] and uses Construction Specifications Institute (CSI) MasterFormat to manage all cost item data for all types of construction, such as steel, concrete, wood, and so forth. It should be noted that the proposed method is also applicable to cost items in other sources. This research mainly focuses on the wood and concrete work, and the case study in this research is a wood framing building with a concrete basement. Therefore, typical cost items for wood buildings are selected in the data collection.

Figure 2 shows several examples of RSMeans cost items. As shown in Figure 2, each work package has unique line lumber and work description. Intuitively, the textual description of work packages is structured so that it should be much easier to extract the desired information from the work description than natural language statements. However, the descriptions of cost items are unstructured data. For example, the item "061110182680, *Wood framing, joists, 2″ × 6″*" is under the item category "Joist framing", which is a sub-category of "Framing with Dimensional, Engineered or Composite Lumber". Ideally, the textual description of item "061110182680" should contain information related to its categories. However, the item "061110182680" is described as "*Wood framing, joists, 2″ × 6″*", while "Framing with Dimensional, Engineered or Composite Lumber" is missing from its work description. In addition, the cost items within the same category have different description

patterns. For example, the items of "061110182680" and "061110182700" are from the same category, "Joist framing". However, the item "061110182680" is described as "*Wood framing, joists, 2" × 6""* while the item "061110182700" is described as "*2" × 8"" wood joist, framing*". Their work descriptions are expressed in a different pattern. As such, the NLP-NER is required to analyze such textual data.

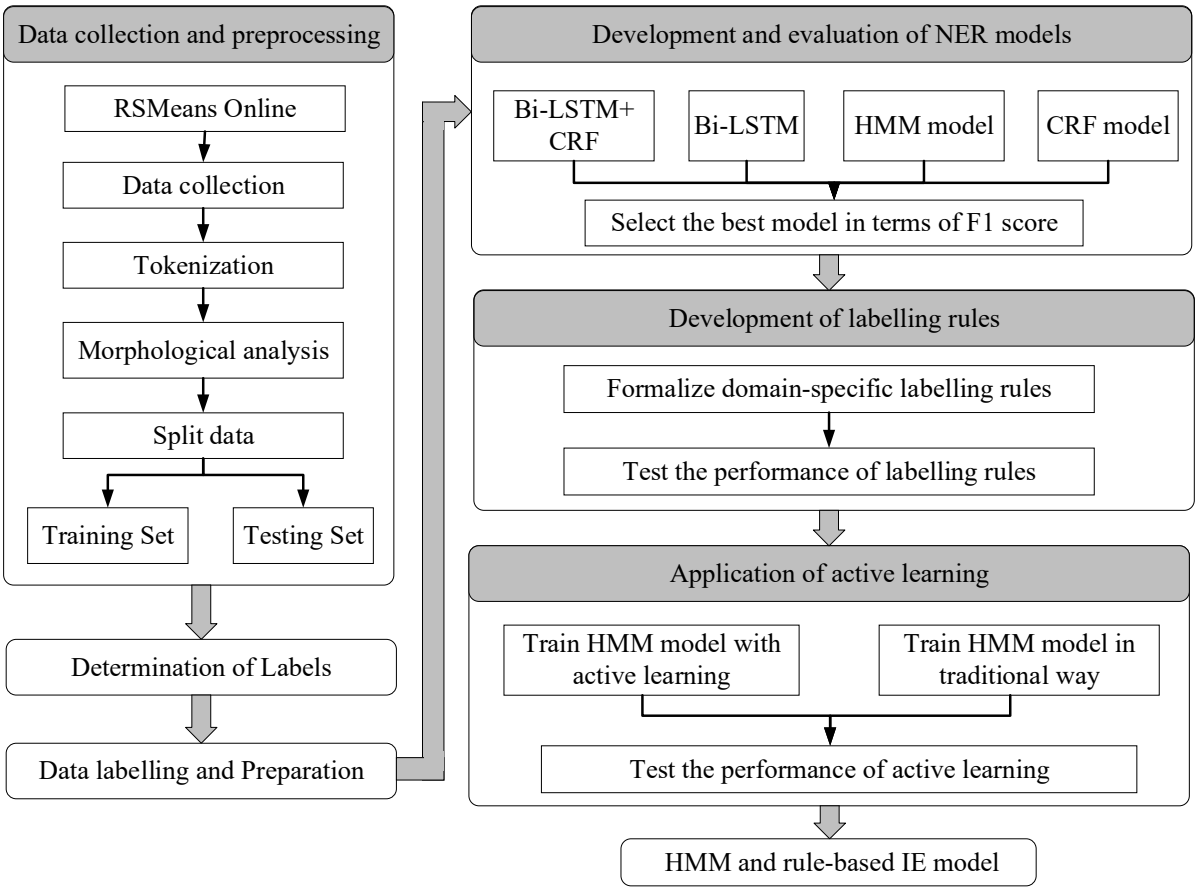

Figure 1. Research Methodology.

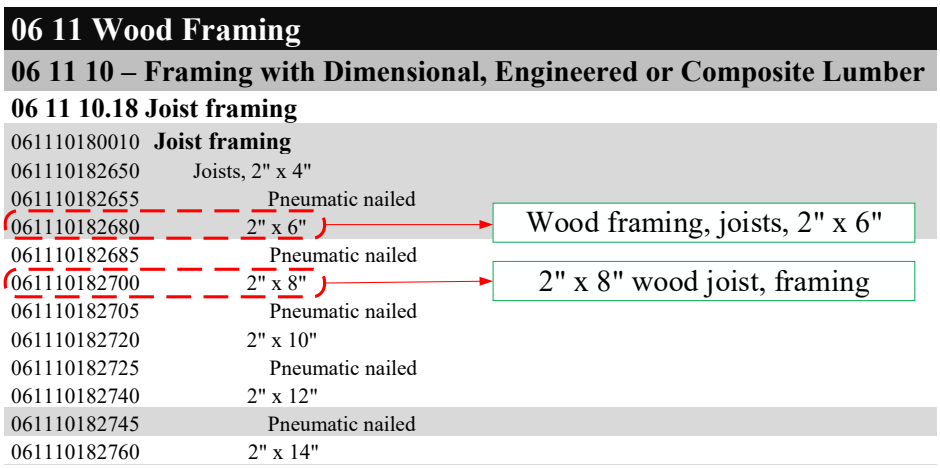

Figure 2. Work descriptions (adapted from RSMeans online [2]).

In the initial stage of IE from work descriptions, the preprocessing is conducted to transform the obtained text data into a clean and computer-processable format. Several NLP techniques were employed, such as tokenization and morphological analysis. For example,

tokenization was used to separate the text into several tokens for feature representation. Generally, the 'word' is a chunk of alphabetical characters separated by space marks, and it is the most commonly used unit in text analysis. Punctuation was also regarded as a token to separate sentences in this study. Following this, morphological analysis (MA) was conducted to identify the different forms of a word and map it to its standard form. MA converts various nonstandard forms of a word (e.g., plural form of the noun) to its lexical form (e.g., the singular form of the noun). Figure 3 presents one illustrative example of text preprocessing. For example, "frames" and "framing" are all mapped to their lexical form "frame", as shown in Figure 3.

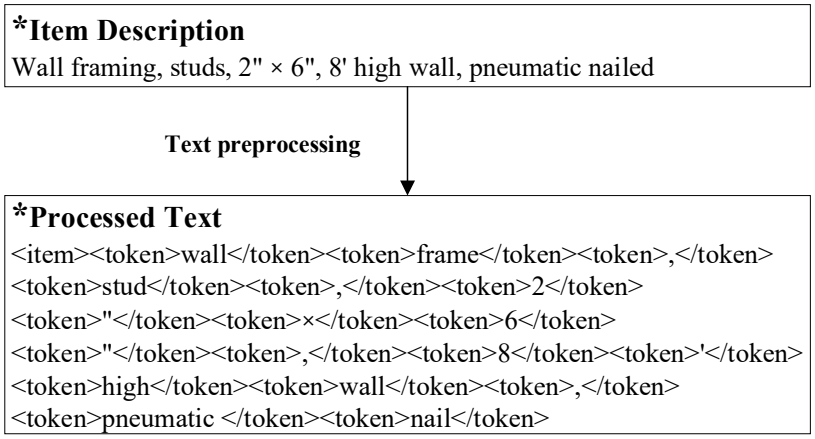

Figure 3. An illustrative example for text preprocessing.

3.2. Identification of Predefined Labels

This research addresses information extraction as a sequence labeling problem through NER models. That is, cost parameters are extracted by assigning their proper predefined labels. The predefined labels of tokens should be determined based on the specific need of the targeted application. For example, the labels can be defined as organization and person, provided that such information is of particular interest. In this research, the desired information is the construction method and product-related features that could be used to query a given BIM model for quantities. Consequently, labels are defined to describe (1) construction activity, (2) construction material, (3) building component, (4) measurement unit, and (5) additional information (e.g., work scope). The description of each category is summarized in Table 1. For example, such labels as material name, type of building element, type of element part, size of building element, function of building element, and material characteristics are defined to describe product-related features.

Table 1. Predefined labels and their description.

Label	Description
M	Product material name
PUK	Punctuation
TBE	Type of building element
TEP	Type of element part
O	Other
SEP	Size of element part
SBE	Size of building element
SHBE	Shape of building element
FEP	Function of element part
CM	Construction method
DF	Design feature
FBE	Function of building element
MP	Material characteristics

3.3. Data Labeling and Preparation of Training and Testing Data

Typically, the NER technique demands manually annotated data to train the ML-based NER models that could be used to classify and label new data/parameters. As such, this step is to prepare work description data and manually label item descriptions. The retrieved dataset was split into two datasets: a training set (80%) and a testing set (20%). The training set was used to train the developed ML-based NER models, while the testing dataset was used to evaluate the performance of the developed algorithms. Each token in preprocessed data is annotated manually after determining token labels. Figure 4 shows one example of the annotated work description. As shown in Figure 4, "Wall Frame" is annotated as "TBE", "Stud" is labeled as "TEP", "2″ × 6″" is annotated as "SEP", and "8′ high" is labeled as "SBE", and "pneumatic nailed" is given a label of "CM".

Annotated Work Description

<item><TBE>wall</TBE><TBE>frame</TBE><PUK>,</PUK>
<TEP>stud</TEP><PUK>,</PUK><SEP>2</SEP>
<SEP>"</SEP><SEP>×</SEP>6</SEP><SEP>"</SEP>
<PUK>,</PUK><SBE>8</SBE><SBE>'</SBE><SBE>high</SBE>
<SBE>wall</SBE><PUK>,</PUK><CM>pneumatic</CM>
<CM>nail</CM></item>

Figure 4. Example of annotated cost item.

3.4. Development and Evaluation of NER Models

Several NER algorithms have been proven to be effective for various construction applications, including (1) Hidden Markov model (HMM) [30], (2) Conditional Random Field (CRF) [31], (3) Bidirectional-Long Short-Term Memory (Bi-LSTM) [32], and (4) Bi-LSTM+CRF [33]. However, there is no evidence indicating that one of them outperforms than others. These algorithms were adopted in this research and were trained based on the retrieved data. As shown in Figure 5, the NER models take a sequence of tokens as inputs and predict corresponding labels. For example, "8′ high" are labeled by NER models as "SBE", "SBE", and "SBE". The conceptual labeling processes of these four models are briefly shown in Figure 5. As shown in Figure 5, HMM and CRF models label every token independently; on the contrary, Bi-LSTM and Bi-LSTM+CRF models can classify all the tokens in a sequence as a whole, capitalizing on the Recurrent Neural Network (RNN).

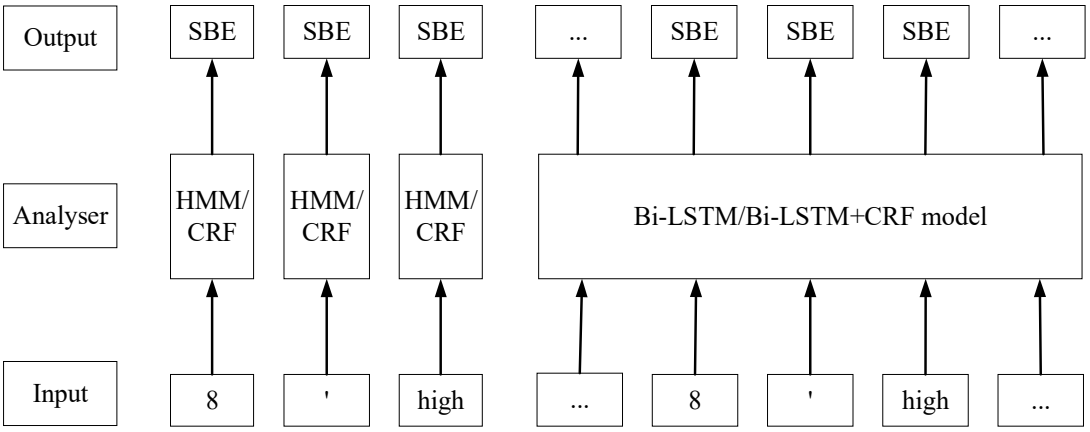

Figure 5. NER-based information extraction.

3.4.1. Feature Engineering

This study employed different strategies of feature representation for each NER model. The HMM algorithm uses the original word to represent each token in a sentence due to its simplicity. In terms of the CRF algorithm, the authors proposed a new feature

representation, i.e., adding syntactic features to express each token in a sentence. A context window of size one is used to capture features of its surrounding tokens to enrich the information contained in the token feature. Figure 6 illustrates the feature vectors for the CRF model. As shown in Figure 6, the generated d feature is a 1 × 12 feature vector. The added features consist of syntactic and semantic features. The syntactic features contain "isDigital" and "isPunctuation". The semantic feature is "isUnit" and can be recognized based on the developed dictionary by comparing the token with each word in the dictionary. The common units in the domain of construction are included in the Unit Dictionary, such as "'", "S.Y.", "HP", and "ga". The token "high" in item description "*Wall framing, studs, 2″ × 4″, 8′ high wall*" is transformed into vector ['', 0, 0, 1, high, 0, 0, 0, wall, 0, 0, 0]. The feature vector indicates that the current token is "high". Its preceding token is "'", while its succeeding token is "high". Moreover, its previous token is a unit.

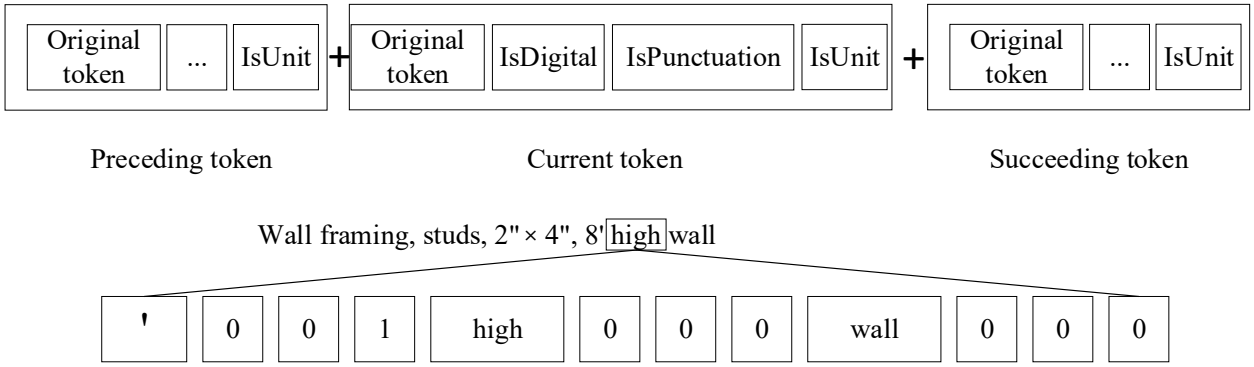

Figure 6. Feature vector for the CRF model.

As for the BiLSTM and BiLSTM+CRF models, the authors employed word embedding to transform textual descriptions of cost items into numerical data so that these two models could take the whole description as the input. The Word2vec method represents a token as a numerical vector, assuming that the meaning of a token can be inferred by its neighbors. Since a piece of item description is a sequence of tokens, a line item is consequently represented as a matrix by the word embedding method. The matrix size is L × d, where L is the number of tokens in the item description and d is the length of a token vector. The word embedding is presented in Figure 7. As shown in the figure, the token "high" was processed in the embedding layer before being fed into the NER model. The skip-gram Word2vec model is a two-layer ANN (Artificial Neural Network). This simple ANN model takes a token as input and returns surrounding words of the target word. After training this model with all samples, weights of the hidden layer have fitted with the training data so that the trained model can predict the context of a given token in samples. Therefore, the hidden layer of the trained ANN model was employed to represent the input token. In this study, the dimension of word vectors, which is also the hidden unit of the skip gram's layer, was set as 140. In addition, the maximum length of the padded sequences was specified as 60 in Bi-LSTM/Bi-LSTM+CRF model, considering the max size of samples in the collected data.

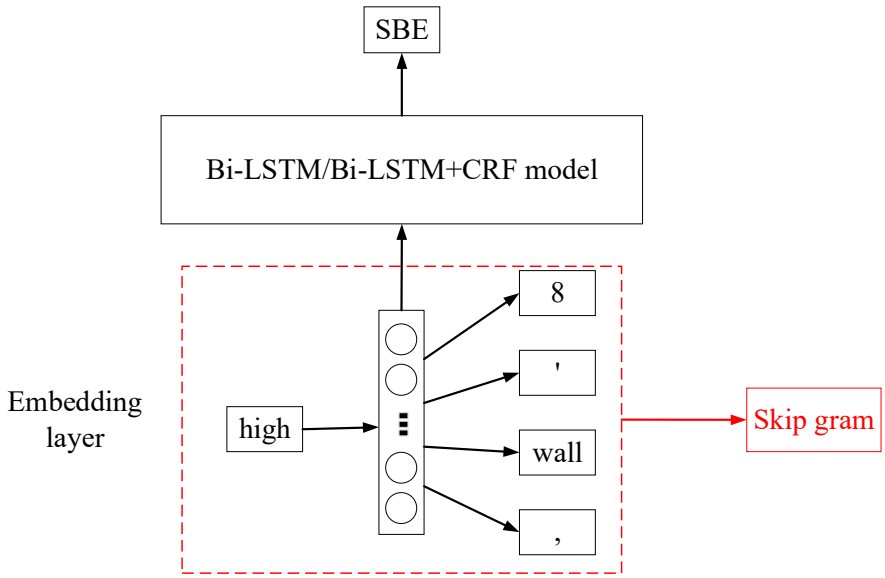

Figure 7. Word embedding for the Bi-LSTM/Bi-LSTM+CRF model.

3.4.2. Model Development

The algorithm selected for HMM model is Viterbi algorithm due to its efficiency in decoding the NER label state sequences [34]. The algorithm in the CRF model is determined as 'lbfgs', which is the abbreviation of "Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm" because of its significant outperformance than GIS and other gradient-based algorithms regarding convergence rate [35]. Then, the number of max iterations was set as 100. The BiLSTM model consists of four layers: (1) embedding layer, (2) bidirectional LSTM (also known as the first LSTM layer), (3) LSTM layer, and (4) time-distributed layer. The embedding layer has been described thoroughly in the last section. The maximum length of the padded sequences was specified as 60 in Bi-LSTM/Bi-LSTM+CRF model, considering the max length of samples in collected data. In training the BiLSTM model, the batch size is set as two because of the limited training data size. In addition, verbose and epochs are determined to 1 and 1 after tuning hyperparameters. Bi-LSTM+CRF model adds a CRF layer to the bidirectional LSTM model. Therefore, the detailed configuration is the same as the previous two models.

3.4.3. Model Evaluation

Three performance metrics are widely used to evaluate the NER models [36], including (1) precision, (2) recall, and (3) F1 score. Precision refers to the ratio of the number of correctly labeled tokens over the total number of tokens. The recall is calculated by dividing the number of correctly labeled tokens by the number of tokens with the same label in the ground truth dataset. On the contrary, the F1 score is the harmonic mean of precision and recall so that it outperforms others in terms of imbalanced class distribution [36]. It is used to measure the performance of the proposed NER-based methods. The performance of these four algorithms is discussed later in the "Case Study" section.

3.5. Sequence Labeling Rules Based on Expert Knowledge

As described earlier, two approaches are typically used for information extraction: (1) ruled-based methods and (2) ML-based methods [11]. ML-based methods eliminate manual effort in rule development by automatically extracting implicit rules (i.e., training the model) from training data. However, ML-based methods suffer from labeling training data and lower prediction accuracy [11,37]. The sequence labeling rules were developed by authors and integrated with ML algorithms described in the previous sub-sections. The formalized sequence labeling rules are summarized in Table 2. For example, rule #1 specifies that when the current token is labeled as "SBE"/"SHBE"/"FBE", its preceding

tokens should contain "TBE" tokens. The entities of "SBE" or "SHBE" provides additional information for tokens with "TBE". For example, detailed information of building elements is always specified after the tokens "Type of Building Element". Similarly, the authors formalized rule #2, because information of "Design feature" is used to describe the building element or element part. Rule #3 indicates that information of "SBE"/"SEP"/"DF" usually contains a cardinal digit. Furthermore, Rule #4 shows that a cardinal digit followed by a unit token is an entity of SBE, SEP, or DF. For example, "2" × 4"" in item description "*Wall framing, studs, 2" × 4", 8' high wall*" is labeled as "SEP", while it contains two cardinal digits and two-unit tokens. It should be noted that the developed rules served as a checker on the results obtained by ML models and cannot work independently.

Table 2. Formalized sequence labeling rules.

ID	Rule
1	The preceding tokens before the "SHBE/SBE/FBE" token should contain the "TBE" token.
2	The preceding tokens before the "DF" token should contain the "TEP" or "TBE" token.
3	Information of SBE, SEP, or DF should contain the cardinal digit.
4	The cardinal digit and its following unit token should be labeled as SBE, SEP, or DF.

Figure 8 provides one example of how the sequence labeling rules work with ML-based NER algorithms. Generally, the ML-based NER model returns a sequence of predicted probabilities of different labels, and the predicted label is the one with the highest probability. Afterward, the developed sequence labeling rules are applied to check the prediction results. As shown in Figure 8, "8'" is labeled as "FBE", which conflicts with sequencing rules. Then, the labeling result with a lower probability will be checked until the proposed labeling rules are satisfied.

3.6. Active Learning

ML-based NER requires human annotators to label a large amount of training data. Such labeling is exceptionally costly and time-consuming [23]. To address such limitations, this study employed the strategy of active learning to minimize the volume of training data, thereby reducing manual labeling efforts. Active learning is to select and learn the most informative-to-learn instances to reduce labeling efforts. In other words, active learning intervenes the selection of the training data for the developed NER model to increase the overall efficiency. Essentially, active learning allows the selection of the most valuable data as input of ML algorithms. For example, '*10" × 10" wood column framing, heavy mill timber, structural grade, 1500f*' and '*12" × 12" wood column framing, heavy mill timber, structural grade, 1500f*' are similar cost items [2]. If the former has been used in the training set, the trained model is expected to label the latter accurately. Active learning, thus, will not feed these two similar items into the training set, thereby reducing the cost spent on labeling data. Figure 9 depicts how active learning works in the proposed NER-based framework. In the traditional method, all the training data are labeled by the human annotator and fed into NER models. In contrast, the training data are inputted to the active learner before being labeled in the scenario of active learning. The active learner evaluates and sorts these unlabeled data in terms of their impact on model training so that the most valuable data selected by the active learner can be fed into the NER model. As a result, there is no need for the human labeler to annotate invaluable data, thereby reducing the manual efforts to label training data. It is important to note that active learning is not intended to improve the performance of the model. It is used to reduce the manual efforts in preparing the training data. Its strategy is to enable the employed model to reach the best performance with minimal training data in the most efficient manner.

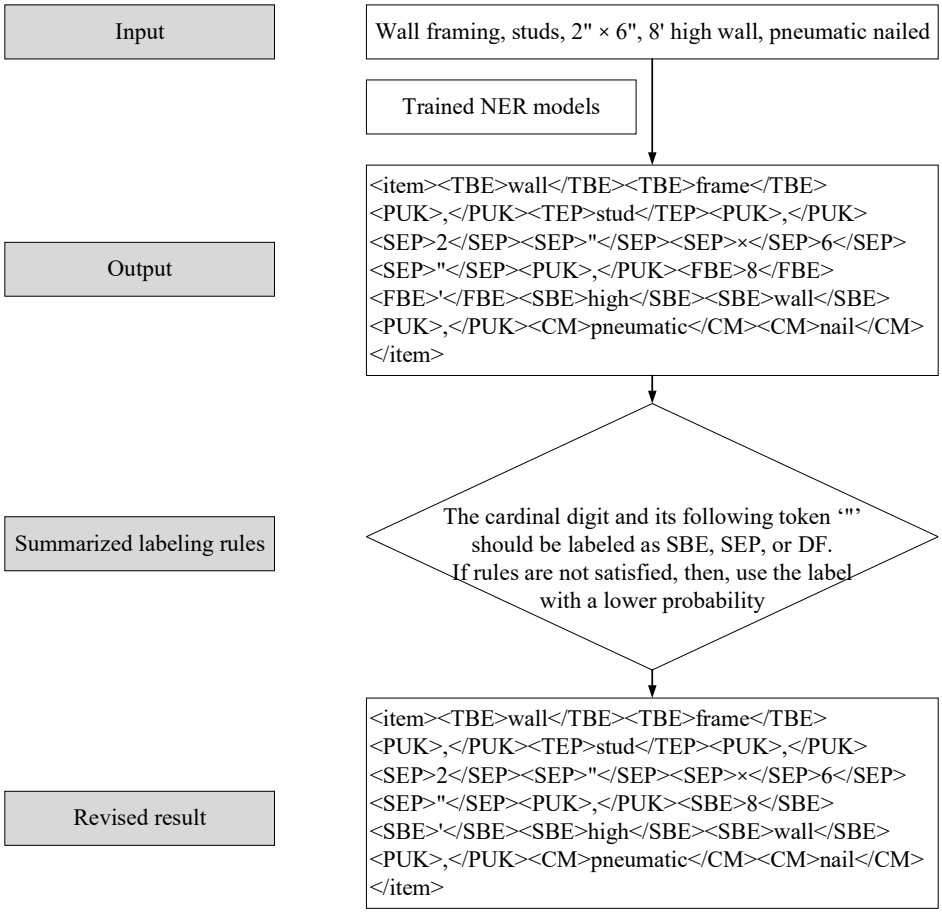

Figure 8. Flowchart and example of sequencing labeling rules with ML-based NER.

Uncertainty sampling is a strategy for identifying unlabeled items that the developed ML model cannot predict confidently. It means that only items with low certainty are selected as the training data. It was adopted by the authors because the developed NER returns the sequence of probabilities that could be used to calculate the uncertainty of the prediction. Figure 10 displays the workflow of the uncertainty sampling method, which functions as an active learner to select training data. Initially, a small portion of the training data is randomly selected from the training data and labeled by the human annotator. Then, these labeled data were fed into the developed NER model. Subsequently, the trained NER model was employed to predict the rest of the training data. Afterward, the uncertainty sampling method is utilized to calculate the uncertainty of the rest of the training samples based on the NER model prediction results. Eventually, the samples with the highest uncertainty are selected as the training data. The prediction uncertainty of the NER model on a piece of work description is quantified by counting the arithmetic average of the uncertainty of every token in the work description of a cost item, employing Equation (1) [23]. The entropy for each token (i.e., uncertainty) is measured using Equation (2).

$$H(item) = \frac{1}{N} \sum_{i=1,2,...N} H(token_i) \quad (1)$$

$$H(token) = -\sum_{i=1,2,...,14} P(l_i|token) log P(l_i|token) \quad (2)$$

where $H(item)$ denotes the entropy of cost item description, $H(token)$ represents the entropy of tokens, N denotes the number of total tokens in a given cost item, and l_i is the categorical label of the token.

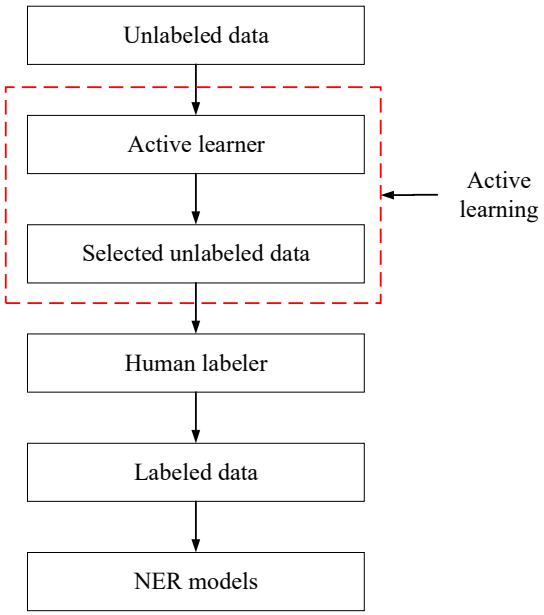

Figure 9. Workflow of active learning.

To test the performance of active learning, two NER models were trained in different strategies. The first one was implemented in the traditional environment. The second one employed active learning. A comparative analysis between two experiments was conducted to quantify the performance of active learning on the developed NER method and presented in the "Case Study" section. In the evaluation of the performance of the active learning method, the involved labeling effort is an important aspect. In this research, the manual annotation effort spent on every cost item is assumed to be equivalent. The human labeling effort is measured based on the size of the required training data.

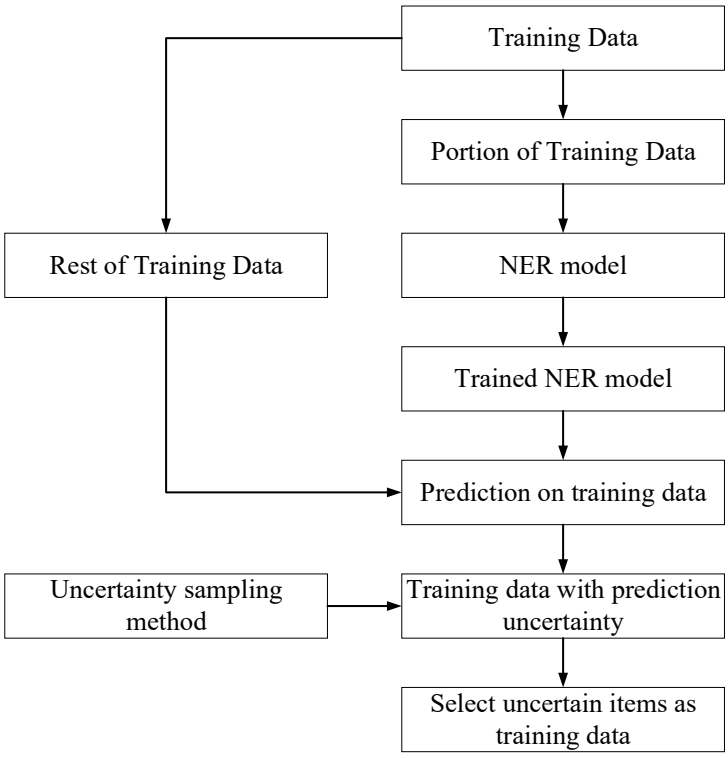

Figure 10. Uncertainty sampling method.

4. Validation

The proposed approach was implemented using the Python programming language. This is because Python is one of the most commonly used ones and is characterized by being open-source, with flexible syntax and good extensibility [38]. In addition, several libraries have been compiled in support of Python. Among these libraries, NLTK is employed to conduct tokenization for the item description and morphological analysis for each token, while pandas are employed to collect data through reading excels files containing a textual description of cost items. Furthermore, the development of HMM model requires sklearn-learn, while CRF model is established with the assistance of Sklearn_crfsuite. Additionally, TensorFlow is utilized to build Bi-LSTM and Bi-LSTM+CRF models.

4.1. Case Studies

RSMeans online cost database consists of thousands of cost items, which cover all types of construction projects. The present research primarily focuses on building construction with a particular focus on light-framed buildings. This type of building, in general, is made of wood-framed superstructures and a concrete basement. Therefore, cost items related to 'Concrete' and 'Wood, Plastic and Composites' are selected to test the proposed NER-based framework. Eighty-three cost items were selected from these two categories and saved into an Excel sheet. Among these extracted line items, 52 items are under the category of "Wood, plastic and composites", while 31 items come from the category of "Concrete". Table 3 shows several examples of these cost items. Their line number, textual description, required crew, and quantity unit are provided in Table 3. These items are selected as representative items, as they are commonly used in the estimation of light-frame buildings.

Table 3. Cost item examples from RSMeans online [2].

Line Number	Description	Crew	Unit
061110307060	Wood framing, roofs, rafters, to 4 in 12 pitch, 2″ × 8″	2 Carp	M.B.F.
061110307000	2″ × 6″ rafters, roof framing, to 4 in 12 pitch	2 Carp	M.B.F.
061110306070	Wood framing, roofs, fascia boards, 2″ × 8″	2 Carp	M.B.F.
061323100500	10″ × 10″ wood column framing, heavy mill timber, structural grade, 1500f	2 Carp	M.B.F.
066310100550	Plastic (PVC) handrails, post base trim, 4 × 4 post	2 Carp	Ea.
061110406140	Wall framing, studs, 2″ × 4″, 8′ high wall	1 Carp	M.B.F.
033113700400	Structural concrete, placing, column, square or round, pumped, 12″ thick, includes leveling (strike off) & consolidation, excludes material	C20	C.Y.

4.2. Results and Discussions

The comparison among different NER models is tabulated in Table 4. As shown in Table 4, the developed models can accurately extract cost parameters from textual information of cost items. The F1 scores of all four NER algorithms are greater than 0.75. However, it is worth noting that these models did not perform well in labeling entities of category "FBE" with an average F1 score of 0.482. Among these four candidates, HMM algorithm achieved the highest average F1 score, i.e., 0.88, which is 5% higher than other NER models. Additionally, HMM model shows the best performance in predicting instances of 8 groups among 13 groups. Consequently, HMM was selected as the NER algorithm. The sample size influences the performance of the Bi-LSTM-based model. However, the HMM model, a stable and straightforward algorithm, works more efficiently with a limited training dataset than deep learning-based models. In addition, the CRF algorithm accommodates context information [39], while HMM algorithm depends only on the previous state regardless of context [30]. Moreover, RNN-based algorithms consider the neighboring words [23]. Traditionally, the consideration of context is an advantage of the NER algorithm. However, the work description of cost items employed in this research is unstructured data. The context of the target token cannot effectively support the NER models to predict its category, resulting in the lower performances of NER algorithms. Figure 11 depicts a labeling example; the work description '*Structural concrete, placing,*

column, square or round, pumped, 12′ *thick, includes leveling (strike off) & consolidation, excludes material*" is tokenized and assigned corresponding labels.

Table 4. F1 scores of four NER algorithms and combined algorithms.

Label	HMM	CRF	Bi-LSTM	Bi-LSTM + CRF	HMM+Rules	HMM + Rules + Active Learning
M	0.95	0.9	0.83	0.85	0.98	0.95
PUK	0.99	0.99	0.95	0.99	0.99	0.99
TBE	0.93	0.82	0.92	0.90	0.95	0.90
TEP	0.57	0.4	0.85	0.8	0.62	0.88
O	0.94	0.81	0.81	0.82	0.98	0.94
SEP	0.89	0.87	0.68	0.56	0.92	0.95
CM	0.88	0.57	0.87	0.96	0.91	0.88
SBE	1	1	0.85	0.79	1	1
MP	1	1	0.78	0.625	1	0.86
DF	1	1	0.72	0.6	1	0.97
FBE	0.428	0.5	0.75	0.25	0.68	0.54
FEP	-	-	-	-	-	-
SHBE	1	1	1	1	1	1
Average	0.88	0.82	0.83	0.76	0.91	0.89

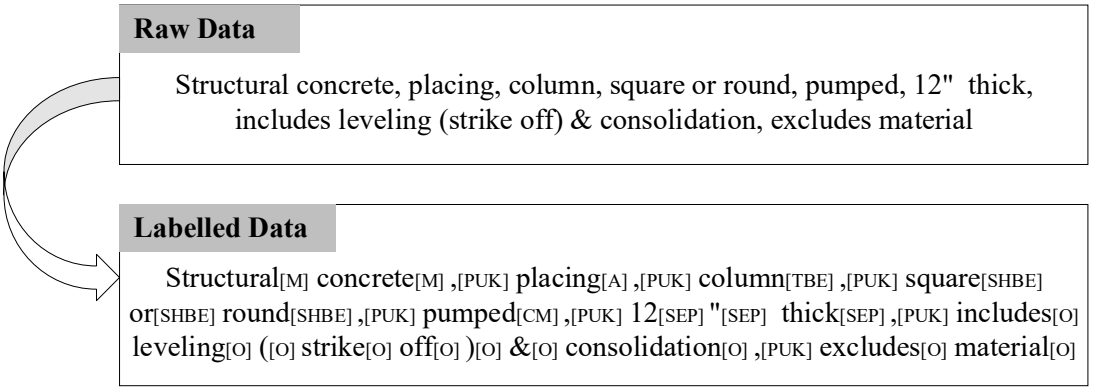

Figure 11. Example of labeled data.

The confusion matrix for the HMM model prediction results is presented in Table 5. The confusion matrix is a contingency table. As shown in Table 5, each column stands for the number of tokens with an actual label, and each row represents predicted instances for each label. The diagonal elements of the confusion matrix are occurrences of tokens that are predicted accurately by the NER model. For example, the middle cell (CM/CM, 32) implies 32 tokens with the actual label "CM" are annotated as "CM" by HMM model. The higher diagonal values of the confusion matrix indicate the better performance of the NER model. The HMM model has poor performance on classifying instances of the 'FBE' category with a 0.428 F1 score. This might be due to the limited size of 'FBE' samples, and most cost items do not contain 'FBE' information.

Two experiments are conducted to test the validity of the developed label rules. The first one employed the HMM model, while the second experiment extracted information by integrating the HMM models and the developed labeling rules. Their detailed performance in terms of F1 score was presented in Table 4. The results revealed that developed labeling rules could generally increase the HMM model. On average, the F1 scores were improved by 3.6%. However, it should be noted that the improvement of the developed rules on the NER model is limited because of two aspects: (1) the HMM model has shown satisfactory performance and (2) a limited number of rules are formalized in this research.

Table 5. Confusion matrix of HMM model.

	M	PUK	TBE	TEP	O	SEP	CM	SBE	PM	DF	FBE	SHBE
M	30	0	0	0	1	0	0	0	0	0	0	0
PUK	0	96	0	0	1	0	0	0	0	0	0	0
TBE	0	0	27	1	2	0	0	0	0	0	0	0
TEP	2	0	0	2	0	0	0	0	0	0	0	0
O	0	0	0	0	138	0	0	0	0	0	2	0
SEP	0	0	0	0	3	33	0	0	0	0	5	0
CM	0	0	0	0	8	0	32	0	0	0	0	0
SBE	0	0	0	0	0	0	0	7	0	0	0	0
PM	0	0	0	0	0	0	0	0	10	0	0	0
DF	0	0	0	0	0	0	0	0	0	3	0	0
FBE	0	0	1	0	0	0	0	0	0	0	3	0
SHBE	0	0	0	0	0	0	0	0	0	0	0	1

In the strategy of active learning, only 49 among 66 cost items are selected as training data for the developed NER model. Table 4 shows the performance of HMM model trained by the active learning method. The result suggested that the trained model with active learning achieved F1 scores of 0.89 slightly lower than 0.91 F1 obtained by the traditional method. Considering that much smaller training data are employed in the active learning method compared to the traditional approach, active learning effectively reduces the manual effort needed to label the text data and train the model. Assuming that the labeling effort is spent on every work item is equivalent, 26% the manual efforts are reduced by active learning. Active learning is used to reduce the manual efforts in data labeling at the expense of reduced performance. Less training data are fed into the developed model due to the employment of active learning; thus, the accuracy and precision of the developed model drop accordingly. However, it is more efficient, since the active learning-assisted model achieved 98% of the performance with 74% of the labeling cost. It indicated that active learning is helpful to find the training data with the most training value.

4.3. Limitations and Future Work

This research has a few limitations. First, the proposed framework was only tested on RSMeans cost items. The cost items from other sources may have a different structure or pattern from RSMeans. The developed method may not work effectively on cost items from other sources. Second, training and testing data are limited. Typically, more extensive data can yield better performance of ML models. However, ML algorithms, including deep learning algorithms, still work well with small data set [40]. There is no explicit definition of the minimum amount of dataset used in machine learning models. It depends on the complexity of the proposed model. The size of the required dataset could be approximately estimated by the first experiment with the created model. Third, in the process of data collection, only cost items within the categories "Concrete" and "Wood, plastic and composites", are considered in the case study. Although active learning is proven to reduce the labeling efforts in the model training, it cannot completely eliminate them. Furthermore, the prediction errors of the proposed automated approach are detected and corrected manually. The automated solution to checking the prediction results will be investigated in the future. In addition, this paper shows the results of the developed NER models for work description analysis, rather than QTO. Estimators essentially need to map the unit price database (providing unit price for cost items) and the BIM models (offering quantities) in cost estimation. Quantities in a given BIM model have to be determined as per the work descriptions of cost items in the unit price database, i.e., mapping quantities and unit prices. The presented research introduced a generic approach for information extraction from work descriptions, laying the foundation for the quantity–price match. The cost items in the proposed approach are not constrained to either a specific unit price database, such as contractors' private price databases or commercial databases or specific annual updates. The authors will employ the extracted information to conduct QTO from BIM design models for automated cost

estimation in the future. Future studies will also be directed at demonstrating the usefulness of the presented research through actual case studies.

5. Conclusions

Intending to automate construction-oriented QTO, this research developed an NER-based information extraction method that extracts information from the work descriptions of cost items. Four NER algorithms, namely (1) HMM, (2) CRF, (3) Bi-LSTM, and (4) Bi-LSTM+CRF, were tested. The results revealed that HMM outperforms others in terms of the F1 score. As such, it was selected to implement the NER-based IE model. Moreover, this study integrated the HMM and manually developed labeling rules. The strategy of active learning was adopted to reduce the number of training data and human labeling efforts. The experimental results showed that the developed NER model (i.e., active learning-based HMM model) could extract the cost parameters from the work item description with satisfactory performance. With the assistance of developed labeling rules, the performance of the ML-based NER model was improved by 3%. The active learning approach could reach the performance of the traditional method with a significantly reduced size of training data, thereby reducing costs for human labeling.

This research contributed to the body of knowledge by the NLP-based IE model integrating HMM and formalized labeling rules that automatically process work descriptions and lay a foundation for automated QTO and cost estimation. This research indicated that HMM algorithm is the most suitable algorithm for IE from the textual description of cost items compared other three common NER algorithms. The integration of HMM and formalized labeling rules has improved accuracy by 89% in NER. In addition, active learning strategies reduced by 26% the labeling efforts for the case study. The represented approach can extract cost parameters from work descriptions, laying a foundation for automated QTO.

Author Contributions: Conceptualization, S.T. and H.L.; methodology, S.T., H.L., M.A., O.A., Z.L., and A.F.; software, S.T.; validation, S.T. and H.L.; formal analysis, S.T. and H.L.; investigation, S.T., H.L., M.A., O.A., Z.L. and A.F.; resources, H.L. and O.A.; data curation, S.T.; writing—original draft preparation, S.T.; writing—review and editing, H.L., M.A., O.A., Z.L. and A.F.; visualization, S.T.; supervision, H.L.; project administration, H.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Peurifoy, R.; Oberlender, G. *Estimating Construction Costs*, 6th ed.; McGraw-Hill Education: New York, NY, USA, 2014.
2. RS Means Data Online. 2021. Available online: https://www.rsmeansonline.com (accessed on 6 December 2021).
3. Monteiro, A.; Martins, J.P. A survey on modeling guidelines for quantity takeoff-oriented BIM-based design. *Autom. Constr.* **2013**, *35*, 238–253. [CrossRef]
4. Zhang, C.; Yao, W.; Yang, Y.; Huang, R.; Mostafavi, A. Semiautomated social media analytics for sensing societal impacts due to community disruptions during disasters. *Comput. Civ. Infrastruct. Eng.* **2020**, *35*, 1331–1348. [CrossRef]
5. Tang, L.; Zhang, Y.; Dai, F.; Yoon, Y.; Song, Y.; Sharma, R.S. Social Media Data Analytics for the U.S. Construction Industry: Preliminary Study on Twitter. *J. Manag. Eng.* **2017**, *33*, 04017038. [CrossRef]
6. Le, T.; Jeong, H.D. NLP-Based Approach to Semantic Classification of Heterogeneous Transportation Asset Data Terminology. *J. Comput. Civ. Eng.* **2017**, *31*, 04017057. [CrossRef]
7. Yu, W.D.; Hsu, J.Y. Content-based text mining technique for retrieval of CAD documents. *Autom. Constr.* **2013**, *31*, 65–74. [CrossRef]

8. Xu, X.; Chen, K.; Cai, H. Automating Utility Permitting within Highway Right-of-Way via a Generic UML/OCL Model and Natural Language Processing. *J. Constr. Eng. Manag.* **2020**, *146*, 04020135. [CrossRef]
9. Zhang, F. A hybrid structured deep neural network with Word2Vec for construction accident causes classification. *Int. J. Constr. Manag.* **2019**, 1–21. [CrossRef]
10. Seedah, D.P.K.; Leite, F. Information Extraction for Freight-Related Natural Language Queries. *Comput. Civ. Eng.* **2015**, *2015*, 667–674. Available online: http://ascelibrary.org/doi/10.1061/9780784479247.083 (accessed on 6 December 2021).
11. Lee, J.; Yi, J.S.; Son, J. Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP. *J. Comput. Civ. Eng.* **2019**, *33*, 04019003. [CrossRef]
12. Zhang, J.; El-Gohary, N.M. Automated Information Transformation for Automated Regulatory Compliance Checking in Construction. *J. Comput. Civ. Eng.* **2015**, *29*, B4015001. [CrossRef]
13. Xu, X.; Cai, H. Ontology and rule-based natural language processing approach for interpreting textual regulations on underground utility infrastructure. *Adv. Eng. Inform.* **2021**, *48*, 101288. [CrossRef]
14. Liu, K.; El-Gohary, N. Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports. *Autom. Constr.* **2017**, *81*, 313–327. [CrossRef]
15. Liu, K.; El-Gohary, N. Semantic Neural Network Ensemble for Automated Dependency Relation Extraction from Bridge Inspection Reports. *J. Comput. Civ. Eng.* **2021**, *35*, 04021007. [CrossRef]
16. Martínez-Rojas, M.; Marín, N.; Vila, M.A. An Approach for the Automatic Classification of Work Descriptions in Construction Projects. *Comput. Civ. Infrastruct. Eng.* **2015**, *30*, 919–934. [CrossRef]
17. Martínez-Rojas, M.; Marín, N.; Miranda, M.A.V. An intelligent system for the acquisition and management of information from bill of quantities in building projects. *Expert Syst. Appl.* **2016**, *63*, 284–294. [CrossRef]
18. Martínez-Rojas, M.; Soto-Hidalgo, J.M.; Marín, N.; Vila, M.A. Using Classification Techniques for Assigning Work Descriptions to Task Groups on the Basis of Construction Vocabulary. *Comput. Civ. Infrastruct. Eng.* **2018**, *33*, 966–981. [CrossRef]
19. Moon, S.; Lee, G.; Chi, S.; Oh, H. Automated Construction Specification Review with Named Entity Recognition Using Natural Language Processing. *J. Constr. Eng. Manag.* **2021**, *147*, 04020147. [CrossRef]
20. Akanbi, T.; Zhang, J. Design information extraction from construction specifications to support cost estimation. *Autom. Constr.* **2021**, *131*, 103835. [CrossRef]
21. Kim, T.; Chi, S. Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry. *J. Constr. Eng. Manag.* **2019**, *145*, 04019004. [CrossRef]
22. Mo, Y.; Zhao, D.; Du, J.; Syal, M.; Aziz, A.; Li, H. Automated staff assignment for building maintenance using natural language processing. *Autom. Constr.* **2020**, *113*, 103150. [CrossRef]
23. Moon, S.; Chung, S.; Chi, S. Bridge Damage Recognition from Inspection Reports Using NER Based on Recurrent Neural Network with Active Learning. *J. Perform. Constr. Facil.* **2020**, *34*, 04020119. [CrossRef]
24. Jallan, Y.; Brogan, E.; Ashuri, B.; Clevenger, C.M. Application of Natural Language Processing and Text Mining to Identify Patterns in Construction-Defect Litigation Cases. *J. Leg. Aff. Disput. Resolut. Eng. Constr.* **2019**, *11*, 04519024. [CrossRef]
25. Akanbi, T.; Zhang, J.; Lee, Y.-C. Computing in Civil Engineering 2019. 2019, no. 2017, pp. 105–113. Available online: http://toc.proceedings.com/49478webtoc.pdf (accessed on 6 December 2021).
26. Lin, J.R.; Hu, Z.Z.; Zhang, J.P.; Yu, F.Q. A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM. *Comput. Civ. Infrastruct. Eng.* **2016**, *31*, 18–33. [CrossRef]
27. Wu, S.; Shen, Q.; Deng, Y.; Cheng, J. Natural-language-based intelligent retrieval engine for BIM object database. *Comput. Ind.* **2019**, *108*, 73–88. [CrossRef]
28. Liu, H.; Cheng, J.C.; Gan, V.J.; Zhou, S. A knowledge model-based BIM framework for automatic code-compliant quantity take-off. *Autom. Constr.* **2022**, *133*, 104024. [CrossRef]
29. News-Record, E. North America's Leading Construction Cost Database. 2019. Available online: https://www.enr.com/articles/48114-north-americas-leading-construction-cost-database (accessed on 6 December 2021).
30. Baum, L.E.; Petrie, T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Ann. Math. Stat.* **1966**, *37*, 1554–1563. [CrossRef]
31. Lafferty, J.; McCallum, A.; Pereira, F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 2001. Available online: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers (accessed on 6 December 2021).
32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
33. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. 2015. Available online: http://arxiv.org/abs/1508.01991 (accessed on 6 December 2021).
34. Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **1967**, *13*, 260–269. [CrossRef]
35. Malouf, R. A comparison of algorithms for maximum entropy parameter estimation. In Proceedings of the 6th Conference on Natural Language Learning, Taipei, Taiwan, 31 August–1 September 2002; pp. 1–7. [CrossRef]
36. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to Roc, Informedness, Markedness & Correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.

37. Zhong, B.; Xing, X.; Luo, H.; Zhou, Q.; Li, H.; Rose, T.; Fang, W. Deep learning-based extraction of construction procedural constraints from construction regulations. *Adv. Eng. Inform.* **2020**, *43*, 101003. [CrossRef]
38. Zou, Y.; Kiviniemi, A.; Jones, S. Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Autom. Constr.* **2017**, *80*, 66–76. [CrossRef]
39. Peng, F.; McCallum, A. Information extraction from research papers using conditional random fields. *Inf. Process. Manag.* **2006**, *42*, 963–979. [CrossRef]
40. Caracol, G.R.; Choi, J.-G.; Park, J.-S.; Son, B.-C.; Jeon, S.-S.; Lee, K.-S.; Shin, Y.S.; Hwang, D.-J. Prediction of Neurological Deterioration of Patients with Mild Traumatic Brain Injury Using Machine Learning. In *Research School on Statistics and Data Science*; Springer: Singapore, 2019; pp. 198–210.