# Extracting Construction Knowledge from Project Schedules Using Natural Language Processing


**Xiaojing Zhao, Ker-Wei Yeoh and David Kim Huat Chua**

**Abstract** A sound and good quality schedule is critical to the success of a construction project. However, the little time available for proper project scheduling in the planning and design stage often impairs the quality of a schedule. Few efforts have been made to evaluate and maintain the schedule quality in the construction stage. Usually project teams need to put intensive manual efforts to conduct schedule quality diagnosis which is time-consuming and subjective to a large extent. One major challenge of diagnosing schedule quality is understanding the activity characteristics and construction logic. The multi-partite nature of construction projects (i.e. schedulers and project teams) further exacerbates the difficulty of diagnosis. This paper thus proposes a novel semantic-based logic reasoning and representation methodology to extract construction methods from the schedule to ensure a consistent project schedule. The intellectual contributions of this paper are twofold. First, this paper develops an ontology of tasks with hierarchies from the schedule to automatically extract the construction methods and activities. Second, this paper presents a novel dependency-based information representation schema for representing the logics between tasks and key constraints to facilitate the complete automation in construction logic reasoning from the schedule. To test the proposed system, this

X. Zhao (✉)
School of Management and Economics, Beijing Institute of Technology, Beijing, China
e-mail: xiaojzhao@bit.edu.cn

Center for Energy and Environmental Policy Research,
Beijing Institute of Technology, Beijing, China

K.-W. Yeoh · D. K. H. Chua
Department of Civil and Environmental Engineering, National University of Singapore, Singapore, Singapore
e-mail: ceeykw@nus.edu.sg

D. K. H. Chua
e-mail: ceedavid@nus.edu.sg



K. Panuwatwanich and C. Ko (eds.), *The 10th International Conference on Engineering, Project, and Production Management*, Lecture Notes in Mechanical Engineering,
https://doi.org/10.1007/978-981-15-1910-9_17

paper evaluates the average rate of recall and precision achieved by the system for extracting construction activities and logics in the schedule within one month and compared the results with the rate achieved by manual check. The developed system provides both academics and practitioners a method to detect the deficiencies of project schedules and assists project planners to produce and maintain good quality schedules starting from project initiation until its completion.

**Keywords** Automatic reasoning · Construction project · Construction knowledge · Ontology learning · Schedule quality

## 1 Introduction

A good project schedule adequately reflects the project scope and defines how and when the project team will deliver the products (PMI 2007). A project schedule with high quality helps improve the planning and control of construction activities and enhances the construction productivity (Bragadin and Kähkönen 2016). Construction activities are subject to many uncertainties, which may lead to multiple schedule disruptions during project execution. Schedule delays may cause a multitude of negative effects on the project. Creating a high quality schedule is a highly complicated task especially in large scale construction projects, as large scale projects are usually characterized of inherent complexity, greater uncertainty and heterogeneous entities with diverse interactions. Despite the importance of project schedules, only a few research efforts have been put forward to examine the quality of schedules.

A number of studies have examined the requirements and performance indicators of schedule quality. U.S. Defense Contract Management Agency (DCMA) (2012) proposed 14 assessment measures for schedule quality control that included logic, leads, lags, relationship types, and critical path check etc. The "cost estimating and assessment guide" report published by United States Government Accountability Office (GAO) provided a best practices checklist for practitioners to manage project cost and schedule (GAO 2009). The schedule quality can be also controlled by the scheduling process (e.g. PMI 2013; Douglas 2006). The process of scheduling should include activity definition, duration estimation, sequencing, resource estimation, schedule development and control (Bragadin, and Kähkönen 2016). A scheduling method prescribes a set of techniques, procedures and rules used by project schedulers (PMI 2007). The scheduling maturity model developed by Association for Project Management and overall quality indicators by PMI can be used to measure the quality of schedule development process (PMI 2013; Douglas 2006).

However, research on schedule quality is still limited. Previous efforts mainly examined the schedule quality from two approaches, i.e. schedule planning, as well as schedule control and evaluation. For schedule planning approach, industrial institutions, such as PMI and DCMA, created standards to define the schedule quality and its development process, and recommended skills and competences required by companies to achieve schedule quality. However, measures to evaluate the schedule at this stage are limited. For schedule control and evaluation approaches, most studies (e.g. APM 2012) examined the compliance of project schedule with predefined schedule assessment criteria based on project schedulers' judgement. Current methods and techniques for schedule checking might be inaccurate and inefficient for large schedules.

This paper develops an automatic project schedule checking (APSC) system to extract construction methods from the project schedule using natural language processing, with the ultimate aim of checking the completeness and accuracy of a construction project schedule. First, an accuracy check and rectification module was developed to check for spelling errors and informal or inconsistent abbreviations in the project schedule; Second, a construction method extraction module was developed by extracting syntactic and semantic features from the schedule. Third, a web-based ontology was developed to represent the properties and hierarchical structure of construction schedules. The developed ontology serves as a means of construction schedule knowledge sharing and reuse. In comparison with traditional schedule checking methods, the APSC system developed in this paper is expected to improve the schedule quality by reducing errors during the schedule checking process.

## 2 Research Background and Knowledge Gaps

This section reviewed literature on: (1) the principles and assessment methods of schedule quality; (2) natural language processing (NLP); and (3) Ontology development.

### *2.1 Assessment of Project Schedule: Principles and Methods*

Many studies on project schedule examined the quality of project schedule from the perspective of contract management and compliance. Russell and Udaipurwala (2000) identified a series of indicators for schedule quality under four groupings: accuracy and completeness; consistency with other planning documents; good practice/workability; benchmarks for control. A good construction project schedule should comply with contract and planning documents, DCMA (2012) formalized 14 check points for schedule health assessment which include logic check, relationship type, float, resources, critical path check and baseline execution index etc. In addition, some schedule protocols provide additional checks for project schedule

quality such as merge points, diverge points, redundant relationships, and out-of-sequence progress. Farzad Moosavi and Moselhi (2014) summarized 48 schedule assessment criteria from different perspectives, including contractual obligation compliance, completeness, the reasonableness of job logic and realist of activity duration. The top ten amongst includes scheduling process, milestones, procurement, Work Breakdown Structure (WBS) and submittal activities etc.

Aside from contract management and compliance, research and industrial practices attempted to control schedule quality during the scheduling process. Various industrial standards and benchmarking schemes have been provided as references for schedulers. PMI (2006) provided practical standards for industry-specific WBS to support the generation of project schedule. PMI (2007) provided a 'schedule model' that represents how and when the team should deliver the pre-defined project scope. American Association of Cost Engineers (AACE) International developed planning and scheduling guidelines for training and professional development (Douglas 2006). The guidelines recommended the roles/responsibilities and skills/knowledge of a scheduling professional in three sections: project planning, schedule development, and schedule management. GAO (2009) developed ten best practices to maintain an integrated network schedule and ten indicators to assess schedule health. NSAI (2009) described the business requirement specification and specification mapping in project schedule and cost performance management. A number of studies also used the above criteria to examine the quality of schedule. Farzad Moosavi and Moselhi (2014) assessed the schedules against industry benchmarks and job logic of three case building projects. Bragadin and Kähkönen (2016) identified five schedule health indicators (i.e. general requirements, process requirements, schedule mechanics requirements, cost and resources requirements, and control process requirements) and evaluated the overall schedule health through a weighting process.

## *2.2 Application of NLP in Information Extraction*

### 2.2.1 Syntactic Information Representation

NLP algorithms were designed to retrieve information from plain text. One common tool used in NLP is Stanford Parser (De Marneffe et al. 2006). It is a probabilistic natural language parser that exploits the grammatical structure of sentences to enable Parts-of-Speech (POS) tagging, chunking, parsing and Stanford dependencies. POS tagging is the process of labelling words or phrases in a text based on the context and definition of words. The results of POS explain how a word is used in a sentence. Words are classified as nouns, pronouns, adjectives, verbs, preposition, conjunctions, etc. (Toutanova et al. 2003).

Chunking is another widely used NLP technique which separates a sentence into phrases (e.g., noun groups, verb groups) rather than single words (Klein and Manning 2003). Libraries such as Spacy and TextBlob can be used to generate phrase out of text.

Parsing analyzes the grammatical structure of a sentence and identifies phrases and their recursive structure. The parse tree illustrates the syntactic relation among the sentence words. Constituency-based parse tree and dependency-based parse tree are the two types of parse trees that are commonly used.

The constituency-based parse trees contain two kinds of nodes: terminal and non-terminal. All interior nodes are non-terminal nodes (e.g., noun/verb phrase) and all leaf nodes are terminal nodes (e.g., noun/verb). The dependency-based parse trees only contain terminal nodes (e.g., noun/verb). A dependency-based parse tree has fewer nodes in a given sentence than a constituency-based parse tree.

### 2.2.2 Semantic Information Representation

The semantic representation is commonly adopted to leverage domain knowledge in the reasoning process in order to address the complex relations involved in a certain domain. This is vital for construction schedule checking since the descriptions of construction activities and tasks is contextual. The semantic representation facilitates computer interpretability, which is essential to facilitate automatic testing and verification of construction schedules.

The Stanford dependencies provide a simple and uniform representation of semantic relations between words. Dependency representations contain around 50 grammatical relations. A grammatical relation holds between the governor (regent/head) and the dependent. For instance, in the statement "The message is sent by the server", the relation is agent while the dependent is server (De Marneffe et al. 2006). This typed dependency means that server performs the action represented by the passive verb sent.

Another technique to detect semantic relations is Semantic Role Labeling (SRL), also called semantic parsing (Gildea and Jurafsky 2002). SRL identifies semantic arguments associated with verbs (predicates) in a sentence and their specific roles. For instance, given a sentence "Install sprinkler pipes", the verb "install" is identified as the predicate, while the message "sprinkler pipes" is identified as the theme. The output is a constituent parse tree that can be transformed into a dependency graph (Björkelund et al. 2009).

In construction applications, Yurchyshyna and Zarli adopted semantic annotation and context-based scheduling to formalize construction conformance requirement in order to realize effective code checking (Yurchyshyna and Zarli 2009). Al Qady and Kandil (2010) utilized shallow parsing to extract semantic knowledge and concept relations from construction contract documents. Two measures, recall and precision, were used to measure the efficiency of Information Retrieval algorithms. Arellano et al. (2015) integrated NLP techniques and application specific ontologies to analyze the requirement specification.

### 2.2.3 Ontology Development

An ontology is defined as an explicit representation of concepts and their relationships in a certain domain. Ontology is commonly developed to provide an information structure and a common understanding for knowledge sharing (Gruber 1995). A number of recent studies used NLP techniques and ontology to extract the knowledge from web pages, and have shown a greater performance in extracting the information. In the construction domain. Creation of OWL ontology not only supports the semantic annotation of text, but facilitates the querying and manipulation of ontology (Zhou and El-Gohary 2017). This study therefore built up a knowledge base for construction schedules, defined all the classes and subclasses with their object properties and relations.

## *2.3 Knowledge Gap*

Despite the achievements mentioned above, most studies on schedule quality assessment manually checked the quality based on predefined criteria and experts' judgement. The schedule management involves daily tasks, duration, location information, resources used and quantities, constraints and milestones. Checking the large amount of records and construction documents manually takes a lot of time and effort. The performance of manual schedule quality check could be subjective and prone to inaccuracy. A few studies recommended the utilization of BIM to automate the generation and update the construction schedule. However, little effort has been done to examine the accuracy and completeness of the project schedule in the project planning stage.

The application of NLP techniques is a promising option that can streamline information extraction and reasoning from construction documents, thereby enabling automatic extraction of construction methods and dependency logic among construction tasks. However, the application of NLP in construction schedule is still in its infancy, and a system is needed to extract the construction knowledge and encode its description and their dependencies.

## 3 Proposed Approach for Construction Knowledge Extraction from Project Schedule

The system architecture of the proposed approach is summarized in Fig. 1. Our approach supports the extraction of both syntactic and semantic structures from the activity descriptions in the project schedule. The proposed system contains three main components: Schedule accuracy check module, Construction method extraction module, and Construction schedule representation module.

### *3.1 Schedule Accuracy Check*

The project schedule document was first pre-processed. This pre-processing phase is called text normalization, which includes the removal of unnecessary marks, punctuations, white spaces, special symbols and stop words. Stop words refer to the words that do not carry important meaning such as "the", "a", "on" and "all". In addition, all the letters were converted to lower case. After pre-processing, the text was tokenized

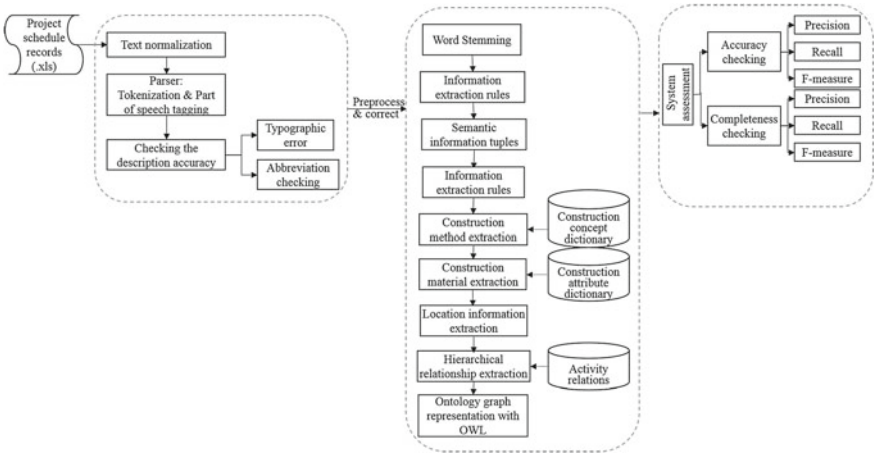


**Fig. 1** The architecture of the proposed APSC system

into sentence, and further parsing and morphological analysis were carried out to bring it into singular form.

The document was then stemmed to reduce words to their root form (e.g., formworks-formwork; painting-paint). Porter stemming algorithm (Porter 1980, 2001) and Paice/Husk stemming algorithm (Zamora 2019) are the two major methods to obtain the word stem. Porter stemmer was applied in this study. It is assumed that there is no stem dictionary and an explicit list of suffixes was given as a criterion to reduce a word to its valid stem.

Three types of errors typically exist in the project schedule, namely, typographic error, cognitive error, and unstandardized abbreviations/descriptions. Typographic errors refer to mistyped words and the correct spellings of the words is known (e.g. peparation-preparation; scuper-scupper). These errors occur when the correct spelling of the word is known but the word is mistyped. Cognitive errors occur when the correct spellings of the word are unknown. The pronunciation of misspelled word is similar to that of the intended correct word. In the case of cognitive errors, the pronunciation of misspelled word is the same or similar to the pronunciation of the intended correct word. The unstandardized abbreviations refer to the abbreviations that are commonly used by construction schedule but cannot be recognized in NLP (e.g. MEZZ; SCDF).

In order to check and modify inaccurate spell error in the activity description, the dictionary lookup and n-gram analysis were used. The error rectification was realized through comparison between misspelled string with the dictionary of words. The word with minimum edit distance was chosen as the correct alternative. These methods can be thought of as calculating a distance between the misspelled word and each word in the dictionary or index (Mishra and Kaur 2013). The shorter the distance the higher the dictionary word is ranked. The interactive spell checking was used to check whether each word is in dictionary and suggested corrections were recommended.

### *3.2 Construction Method Extraction*

Construction methods describe the procedures and techniques that are used in the construction process. In this paper, a construction method is defined as a series of sequential construction activities. A description of construction activity usually consists of construction action, materials or elements, and location information. In a project schedule, a description of construction method is also linked to its duration, start and end date. In order to extract the valid concepts of construction methods from corpus, POS tagging such as noun phrase (NP), verb phrase (VP) and adjective phrase (AP) was first assigned to each word of the rectified schedule text based on its context and definition. A rule-based shallow parser was then used to break the sentence into clauses, and the words in each clause were further tagged into NPs, VPs and APs etc. and its roles (e.g. SUBJ, DOBJ and ACTIVE_VERB). In order to extract the key phases of construction method, the extracted template was set as

JJ-NN, NN-NN (NNP-NNP) and JJ-NN-NN, e.g. "Install-slab-bottom-rebar" (NNP-NN-NN-NN) and "HCS-installation" (NNP-NN). The NLTK package was used to generate POS tagger.

Named Entity Recognition (NER) tool was used to find named entities in text and classify them into pre-defined categories. The extracted phrases formed the activities in construction methods with the help of dictionary. First, construction activities were extracted from the text. Each formed action was extracted based on construction action dictionary. Construction material/elements extractor was designed to identify the material/element (subject or object) related to each action is divided into two categories. After the extraction of construction activities, the hierarchical relationship between activities is further inferred by two different methods: 1. Inherently nested relationships between activities, and 2. Hierarchical relationships defined by start and end dates.

## *3.3 Construction Schedule Representation Module*

The paper presents the construction methods identified from project schedule in RDF/OWL format. A schedule ontology was developed with Protégé software. Figure 2 represents the construction activities in the schedule ontology and a screenshot of the ontology implemented in the software. Developed by the Stanford center for biomedical informatics research, Protégé provides an open-source platform for ontology development to represent the domain knowledge base. As showed in Fig. 2, a construction activity has data type properties that define its action, material/element, location, duration, lag, and predecessor. The hasElement object property relates activities to building material/elements. The dependencies between an activity and other activities are described by its lag and predecessor. The dependency between the concerned activity and predecessor includes finish to finish, finish to start, start to finish, and start to start. The hasLag property describes the number of lag days between two activities. In a construction schedule, an activity may represent the construction of one element or several elements.

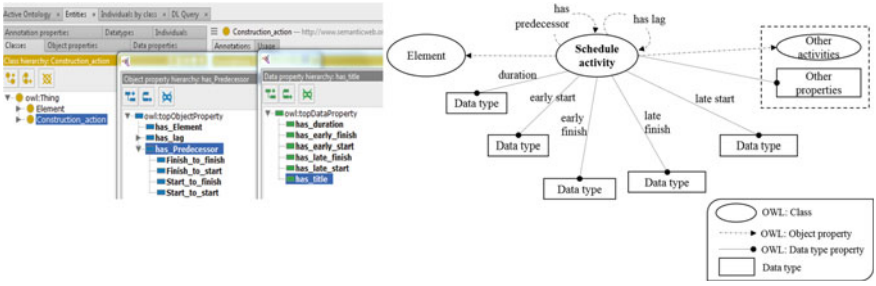


**Fig. 2** The ontology of project schedule

# 4 Implementation of the Proposed System

The proposed system has been implemented using JAVA platform. The Java API for Stanford parser is integrated with the rest of the modules written in Java. Finally, an ontology in OWL format supported by Protégé represents the knowledge base of project schedule.

## *4.1 Schedule Accuracy and Rectification*

The accuracy check results produced by the system for a case project schedule (including activity description) are presented in Fig. 3. The input text is part of activity description in the project schedule. The results of spelling checking not only extracted the spelling errors in the activity description, such as 'scuper', but also the informal/inconsistent abbreviations used by various parties such as 'LT' and 'HT'. After automatic extraction and ratification process, the tree structure generated from shallow parsing in Fig. 4 shows the identified noun phrases of construction methods in the schedule.

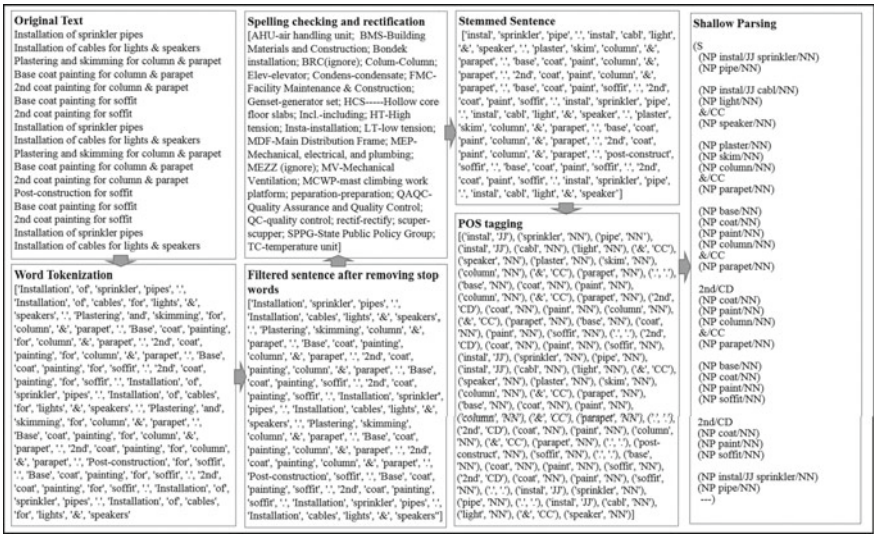


**Fig. 3** An illustration of inputs and outputs of main NLP steps


**Fig. 4** Output from the chunking of schedule description

## *4.2 Schedule Knowledge Base in OWL Format*

The class tree of construction methods generated from given input is then exported on an ontology editor (Fig. 5). An ontology of project schedule in web-based format integrates the NLP results and systems requirements and helps organize the semantic features of project schedule in the hierarchy structure. The developed ontology serves as a proof of construction methods and stores the hierarchy and data in a relational database.

## *4.3 Validation*

For the validation of the developed system, a master plan of a warehouse building project in Singapore is used. The building consists of five storeys with different elements on each floor. Considering this research domain is still a nascent area, the validation is limited to the performance of the system in extracting construction methods from the schedule.

The activity descriptions in project master plan were processed using the NLP framework described in research methodology. The evaluation results of schedule accuracy and completeness are summarized in Table 1. Three simple indicators were used to measure the performance of the system. The performance was evaluated using precision, recall, and F-measure, which combines precision and recall into one measure. Precision rate (P) measures the percentage of correctly extracted activities relative to the total number of activities extracted. Recall rate (R) measures the percentage of correctly extracted activities relative to the total number of activities existing in the source text. F-measure is calculated using Eq. (1) (Toutanova et al. 2003):

$$F = \frac{(\alpha^2 + 1)PR}{\alpha^2 P + R} \tag{1}$$

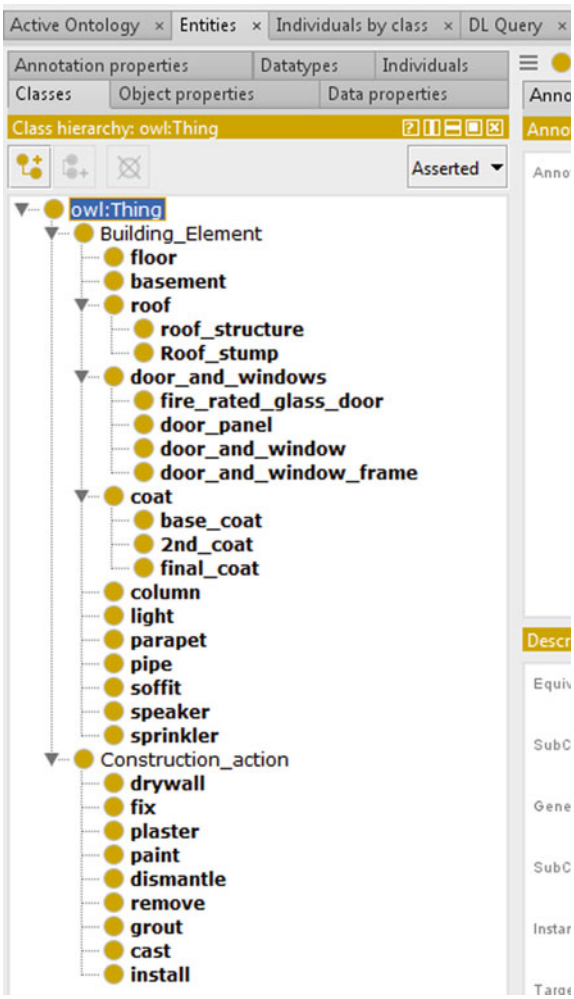


**Fig. 5** Partial view of the construction schedule ontology

which α assigns relative weights to P and R value. $\alpha = 1$ in this study. The recall rate of the APSC system in activity extraction outperformed that in accuracy checking, while the precision rate of the APSC system in accuracy checking outperformed that in activity extraction. The results reflect the trade-off between recall and precision.

**Table 1** Results of the development system in extracting construction methods from project schedule

| Extraction completeness assessment | | Accuracy assessment | |
|---|---|---|---|
| Total No. of activities | 127 | Total No. of inaccurate descriptions/words | 53 |
| Total No. of extracted activities | 122 | Total No. of inaccurate descriptions extracted | 50 |
| No. of correctly extracted activities | 111 | No. of inaccurate descriptions correctly extracted | 48 |
| P measure | 92.1% | P measure | 96.0% |
| R measure | 90.9% | R measure | 90.6% |
| F measure | 91.5% | F measure | 94.3% |

## 5 Conclusions

This paper presents an APSC system for automatically extracting construction methods from project schedules to support the automated schedule quality assessment in construction. The combination of NLP techniques and defined OWL-based ontology models were used to extract both semantic features and syntactic features of construction activities. The developed system contributes to the body of knowledge in four main ways. First, the system allows for detecting and rectifying the inaccuracies in project schedule automatically, which avoids both errors and labor inputs resulting from manual schedule check. Second, the system allows for extracting the domain-specific information of construction methods from complex sentence structures, which save the computational efforts resulting from processing irrelevant text in project schedule. Third, the proposed OWL-based ontology allows for capturing the dependency relations among construction activities. The experimental results show that the proposed system is effective and efficient in evaluating the quality of construction project schedule. Two limitations exist in the study. First, the proposed system addressed most semantic ambiguities, however a number of semantic interpretation issues (e.g. POS tagging ‘construct, NN’) still require the human judgement. Future research is needed to realize the fully-automated way of semantic ambiguity and interpretation issues. Second, the proposed system is developed based on the schedule of building projects, additional effort may be required to extend the ontology for applying the system in a different domain such as infrastructure project.

## References

Al Qady M, Kandil A (2010) Concept relation extraction from construction documents using natural language processing. J Construct Eng Manag 136(3):294–302

Arellano A, Carney E, Austin MA (2015) Natural language processing of textual requirements. In: The tenth international conference on systems, Barcelona, pp 93–97

Association for Project Management (APM) (2012) The scheduling maturity model. APM, Buckinghamshire

Björkelund A, Hafdell L, Nugues P (2009) Multilingual semantic role labeling. In: Proceedings of the thirteenth conference on computational natural language learning: shared task, Association for Computational Linguistics, pp 43–48

Bragadin MA, Kähkönen K (2016) Schedule health assessment of construction projects. Construc Manag Econ 34(12):875–897

Defense Contract Management Agency (DCMA) (2012) Earned value management system (EVMS) program analysis pamphlet (PAP), No. DCMA-EA PAM 200.1, Department of Defense, Washington, DC, USA

De Marneffe MC, MacCartney B, Manning CD et al (2006) Generating typed dependency parses from phrase structure parses. Proc LREC 6:449–454

Douglas EE (2006) Recommended practice No. 14R-90: responsibility and required skills for a project planning and scheduling professional, AACE International, Champaign, IL

Farzad Moosavi S, Moselhi O (2014) Review of detailed schedules in building construction. J Leg Aff Disput Resolut Eng Constr 6(3):05014001

Gildea D, Jurafsky D (2002) Automatic labeling of semantic roles. Comput Linguist 28(3):245–288

Gruber TR (1995) Toward principles for the design of ontologies used for knowledge sharing. Int J Hum Comput Stud 43:907–928

Klein D, Manning CD (2003) Accurate unlexicalized parsing. In: Proceedings of the 41st annual meeting on association for computational linguistics, vol 1, pp 423–430

Mishra R, Kaur N (2013) A survey of spelling error detection and correction techniques. Int J Comput Trends Technol 4(3):372–374

National Standards Authority of Ireland (NSAI) (2009) Project schedule and cost performance management (PSCPM), No. CWA 16022:2009, Brussels

Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137

Porter MF (2001) Snowball: A language for stemming algorithms. Available at: https://snowball.tartarus.org/texts/introduction.html. Accessed on 10 Mar 2019

Project Management Institute (PMI) (2006) Practice standard for work breakdown structures. Project Management Institute, Pennsylvania, USA

Project Management Institute (PMI) (2007) Practice standard for scheduling. Project Management Institute Inc, Newton Square, PA

Project Management Institute (PMI) (2013) A guide to the project management body of knowledge. Project Management Institute Inc, Newton Square, PA

Russell AD, Udaipurwala A (2000) Assessing the quality of a construction schedule. Construction congress VI: building together for a better tomorrow in an increasingly complex world. Orlando, Florida, pp 928–937

Toutanova K, Klein D, Manning CD, Singer Y (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 conference of the north American chapter of the association for computational linguistics on human language technology, vol 1, pp 173–180

United States Government Accountability Office (GAO) (2009) Cost estimating and assessment guide, GAO-9-3SP, U.S. GAO, Washington, DC

Yurchyshyna A, Zarli A (2009) An ontology-based approach for formalisation and semantic organisation of conformance requirements in construction. Autom Construct 18(8):1084–1098
Zamora A (2019) Modifications to the lancaster stemming algorithm. Available at: https://www.scientificpsychic.com/paice/paice.html. Accessed on 21 Jan 2019
Zhou P, El-Gohary N (2017) Ontology-based automated information extraction from building energy conservation codes. Autom Construct 74:103–117