Risk Extraction and Analysis of Technical Specifications Based on Machine-Learning Algorithms for EPC Bid Documents

Minji Park , minjipark@postech.ac.kr
Pohang University of Science and Technology (POSTECH), Pohang, South Korea

Eul-bum Lee, dreblee@postech.ac.kr
Pohang University of Science and Technology (POSTECH), Pohang, South Korea

Seung-yeab Lee, sylee@postech.ac.kr
Pohang University of Science and Technology (POSTECH), Pohang, South Korea

Jong-hyun Kim, jonghyun@wise.co.kr
WISEiTECH co., Seoul, South Korea

Abstract

Most Engineering-Procure-Construction (EPC) companies do not have a system that supports knowledge-based systematic decision-making of technical specifications provided by the contractor, which exposes them to many project risks in bidding or project execution stage. Thus, this study developed two modules for automatic risk extraction and analysis of EPC engineering technical specifications. The first is the technical risk extraction (TRE) module. This technology enables detection and analysis of technical risk clauses that are likely to be missed in bidding stage due to time constraints or lack of personal competence. The second is the Standard Design Parameter (SDP) comparison module. It is a module that allows users to detect design differences or errors by comparing the design standard with the numerical requirements of the technical specifications to be analyzed. Through the above the algorithm models, we implemented a theoretically based system that can be applied to project risk minimization and user collaboration.

Keywords: Technical Specifications, Risk Extraction, Machine Learning Algorithm, Decision Support System, EPC

1 Introduction

EPC projects, which are the background of this study, carry out all the processes from Bid to Operation and Maintenance (O&M). In general, the EPC type of project contract can be said to be a unilaterally advantageous contract type for the companies. In addition, in the case of EPC competitive bidding, it induces company to win orders at less than an appropriate price so that the company can make a significant profit. The most important reason for large losses in EPC project is that contractors receive orders below reasonable prices in competitive bidding. However, for the success of the project (Micheli et al. 2009), it is necessary to minimize contract risk by bidding at a reasonable price, not by analysis to lower the bid price.

Thus, in this study, two algorithm models were developed for automatic risk extraction and analysis of EPC engineering technical specifications as follows. This study is based on structured analysis data on technical specifications collected from the companies. When a technical specification that requires analysis is entered into each module, the following results are verified. The TRE module provides an 'Evaluation Score' for the severity through each provision and the Phase-Matching work with Technical Risk Lexicon. The Evaluation Score enables users to assess the risk of technical specifications for the project based on the severity range determined through the normalization process.

The second SDP comparison module detects the core requirements (Numerical requirements) of the companies in the technical specification. It detects design errors or differences from standards compared to design standards (International Code or Standard). Using the Context-Managing technique, the design criteria are found through context analysis, and the analysis results are presented to users by comparing them with the design standards embedded in the system. As shown above, two algorithm modules developed in this study extract risk clauses in engineering technical specifications and detect technical errors. The TRE module can evaluate project risks at the bidding stage to support management decision making, and the DPE module is expected to be useful for practitioners in analyzing design requirements in the field. As a result, the purpose of this study is to provide a system for preventing risks arising from project bids and execution by EPC contractors.

2 Literature Review

In order to carry out this study, the following academic papers or precedents were reviewed. Based on research on risk model development (Lee et al. 2019) during the EPC bidding phase, it focused on the development of risk automatic extraction model. There is a paper that analyzes and evaluates the risk of the high-level steps to implement the EPC project (Son & Lee 2019). However, at the upstream stage, examples of risk analysis studies were insufficient. Therefore, this study aimed to analyze the technical specifications provided by the companies, especially among the bid document (ITB) during the EPC upstream stage.

Next came the need for the development of machine learning-based decision-making systems (Sackey & Kim 2018) when developing models. Although the use of machine learning technology is a recent trend when analyzing the contractual terms (Fantoni et al. 2021) of a bid document (ITB) provided by the clients, there is a lack of risk extraction research cases targeting technical specifications (Saint 2018). Also, there are studies that suggest algorithms for risk extraction, but only methodology (Putra & Triyono 2015). Although there have been system cases for decision-making in other areas (Zhuang 2021), it is difficult to find system cases associated with technical specifications. Therefore, the decision support system using machine learning (Ferrucci et al. 2010) was applied to the analysis algorithm of technical specifications in this study.

3 Research Process and Methodology

The course of this study is divided into five (5) stages. First, the literature review for the algorithm analysis is conducted, followed by the algorithm development, system application, interpretation and evaluation of analysis results. The course of this study is shown in Figure 1 below.

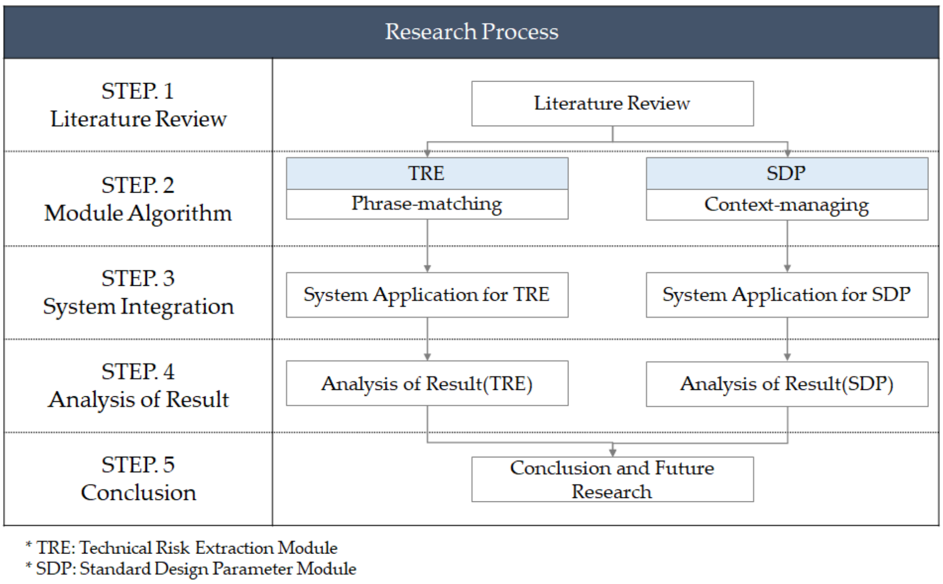

Figure 1. Research process of Technical Specifications Analysis

STEP 1. This is the stage of conducting a prior study for the target setting of this study. Through this process, we suggest the need for automatic risk extraction capabilities for technical specifications at the EPC bidding stage.

STEP 2. This step is the algorithmic construction phase. This includes data collection in technical specifications, data formalization, and the algorithm development. Technical specifications for EPC projects were collected from the companies. Based on the collected technical specifications, the TRE module's Technical Risk Lexicon (TRL) was established, and the SDP module's Standard Design Parameter (SDP) and Synonym Dictionary were established. Machine learning techniques are applied to the construction of the algorithm logic. In the TRE module, we developed an algorithm that can analyze risk severity through Phrase-Matching technology of Natural Language Processing (NLP) even among machine learning. In the SDP module, we applied algorithms that can be derived from the results by comparing the design standard (International Code or Standard) with the design parameter in the corresponding technical specification through the context-managing technology. Details are described in Chapter 4, Module Algorithm.

STEP 3. In this stage, two algorithmic modules were implemented in the system. The two modules were implemented in dashboard format on the system platform to present analysis results. The program user selects the required function out of modules when analyzing technical specifications. The technical specifications were uploaded and analyzed. The results of the performance can be checked visually on the screen. Details are described in Chapter 5, System Integration.

STEP 4. System users can download and utilize analysis results. The results of the system's analysis support decision making to users. The analysis results of the TRE module indicate the evaluation results based on the severity of the technical specifications. The SDP module presents a comparative analysis result table of the design requirements of the technical specifications. The details are described in Chapter 6, Analysis of Result.

STEP 5. As a conclusion step, quantitative evaluation of the output was carried out through analysis results using the above two modules. Also, we propose the research direction of the module for future research of technical specifications. Also, the details are described in Chapter 7, which makes up the Validation & Conclusion parts.

4 Module Algorithm

It describes the algorithm construction process of the TRE module and SDP module. As shown in Figure 2, the two algorithms require embedded data. In other words, it is the baseline data of the system that is utilized to perform the analysis. The data in the technical specifications to be

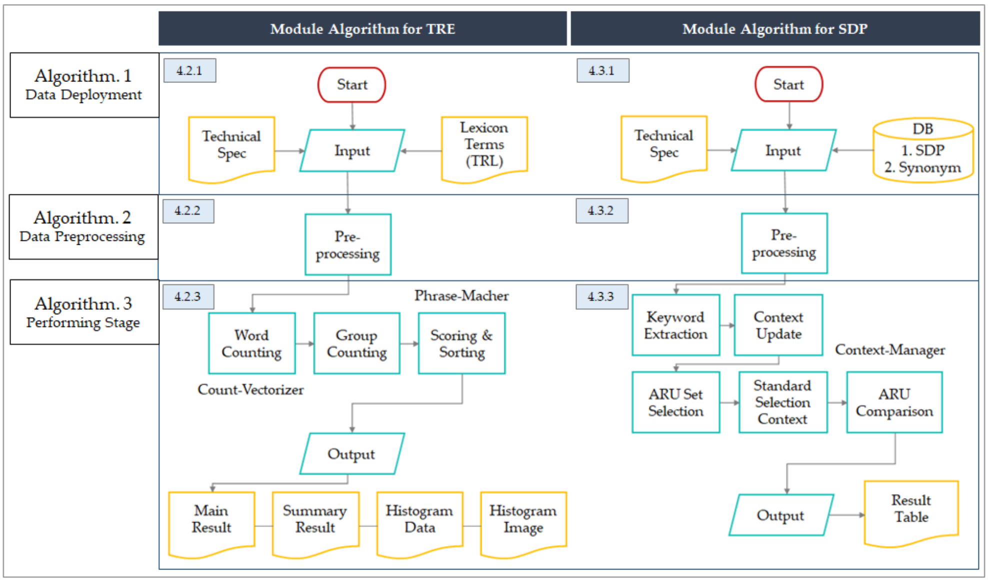

Figure 2. Overall algorithm flow chart for the two modules

analyzed in response are new data. We looked at the process by which baseline data and new data are used within the algorithm.

4.1 Algorithm Structure

The overall algorithm flow chart for the two modules is the same as Figure 2. The algorithm is divided into three stages. It is carried out in the order of the stage of building baseline data, the stage of preprocessing new data, and the stage of the algorithm analysis.

4.2 Algorithm Logic for TRE

This module is an algorithm for extracting technical risk phrases within technical specifications based on machine learning. We have established a baseline data (Technical Risk Lexicon) required for analysis. The phrase-matching technique of NLP was used to extract risk keywords within technical specifications.

4.2.1 Data Deployment

Based on fifteen (15) technical specifications collected from EPC companies, TRL was established. The TRL was finalized through consultation and review by Subject Matter Experts (SMEs) with 15 to 20 years of experience in extracting case-based project risk keywords and performing EPC projects. TRL is a Lexicon for risk terms for extracting risk keywords. As shown in Table 1, three groups were grouped according to keyword severity.

Table 1. Technical Risk Lexicon (TRL)

Clause Type		
High Impact/ High Probability (3)	**High(Medium) Impact/ Medium(High) Probability (2)**	**Medium Impact/ Medium Probability (1)**
all	unless otherwise specified	in compliance with
throughout	unless otherwise mentioned	in accordance with
owner	unless directed otherwise mentioned	according to
company	approved by	shall comply with
no additional cost	not exceed	shall submit
by the bidder	not applicable	discrepancy
by the contractor	not permitted	still
contractor shall include	not allowed	even
⋮	⋮	⋮

To prepare the TRL, case-based risk keywords by Discipline were classified based on technical specifications. Approximately 450 risk clauses were collected for each discipline type, such as mechanical, electrical, instrument, civil, construction, firefighting, HVAC, and piping. These keywords were classified into three groups according to Impact and Probability in an advisory team of five (5) SMEs. Risk clauses, Upstream & Downstream Engineering Process, etc. were comprehensively determined and selected.

4.2.2 Data Preprocessing

Data Pre-Processing is the process of converting unstructured data into structured data. This is an essential prerequisite for performing two modules. This is because all documents should be classified into data that computers can analyze. Text data is distinguished through Sentence Tokenizer. When the unit of a token is a sentence, this task is to separate the text into sentences within the corpus. If sentence classification is made based on the punctuation mark of a sentence using the usual Sentence Tokenizer method, it is not suitable for sentence classification in a technical specification with multiple punctuation marks expressing the position number of a sentence, such as "4.2.2". Therefore, in this study, sentence formats in technical specifications were reviewed among ITB documents (Oevermann 2018), and rules were defined directly according to how Punctuation marks and special characters were used.

4.2.3 Performing Stage

The TRE algorithm consists of three stages. It proceeds to Corpus count Vectorising, Phrase matching, and output (data frame) formation. First of all, Word Counting is performed on the Corpus count Vectorising stage. In other words, it is a step in counting the vector values of the risk keyword. The code function uses Count-Vectorizer to tokenize the sentence and count the number of each token. Next, the Phrase Matching step is the process of extracting word frequencies from vector values and grouping them into groups. The total score of the risk sentence is determined by the score of each group. According to the score, the total score of the risk sentence is sorted in order of highest. The final output formation phase provides four (4) analysis results and a summary table. A detailed analysis of the outputs is described in Chapter 6, Analysis of Results.

4.3 Algorithm Logic for SDP

This module performs a comparison with the design parameters of the target to be compared with the equipment specific SDP that is embedded in the system. Through this process, the conditions and scope of the technical specifications required by the companies are analyzed and the comparative results are presented.

4.3.1 Data Deployment

The embedded data in this module consist of two types. The first data is the Standard Design Parameter (SDP) by the equipment such as vessel and instrument. SDP is the standard data for comparative analysis, as shown in Table 2. The second data is Synonym Dictionary, and the frame is as shown in Table 3. Parameters of SDP were sometimes expressed differently within the same technical specification. There were also cases where one parameter was written in the same meaning but different expressions in each technical specification. Therefore, Synonym Dictionary is constructed to enhance the accuracy and reliability of the analysis. The two embedded data are utilized in the context-managing analysis of the algorithm.

Table 2. Standard Design Parameter (SDP) of Pressure Vessels in Mechanical process (example)

Discipline	Equipment	PRM1 Definition	PRM2 Component	PRM3 Sub element	Attribute	Range 1	Range 2	Unit
Mechanical	Pressure Vessel	Design Temperature	Service	Hydro carbon	greater than	120		° C
Mechanical	Pressure Vessel	Design Temperature	Service	Chemical	greater than	120		° C
Mechanical	Pressure Vessel	Design Temperature	Service	Steam	greater than	120		° C
Mechanical	Pressure Vessel	Design Temperature	Service	Wet sour	max	200		° C
Mechanical	Pressure Vessel	Design Pressure	Service	Hydro carbon	greater than	1.7		MPa
Mechanical	Pressure Vessel	Design Pressure	Service	Chemical	greater than	1.7		MPa
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3. Synonym Dictionary of SDP (example)

SYM Category	STD Word	SYM Word
P(Parameter)	Design Metal Temperature	MDMT
P	Carbon steel/LAS	Carbon Steel/Low Alloy Steel
P	Stainless Steel	STS
P	Minimum Thickness	smallest thickness
P	Design Temperature	Design temp

A(Attribute)	to	~
A	max	≤
A	max	maximum
A	max	not exceed
A	max	not be greater than
U(Unit)	%	percent
U	° C	degree C
U	dBA	dB(A)
U	in	inch
⋮	⋮	⋮

4.3.2 Data Preprocessing

This is the internal pre-process of the SDP module. To select SDP parameters in text, the user selects the type of discipline and equipment. The process of removing the remaining text, leaving only the SDP data corresponding to the selected discipline and equipment. Next, re-process SDP data and Synonym data using dictionary among Python's data types. This reduces the time when searching for SDP or Synonym parameters in the text on code. Next, we analyzed the parameters by converting all into plural and adding them to the Synonym dictionary. Finally, through the tokenization process, we generated an N-gram that sees N phrases grouped by N tokens as a unit. The above processes conclude the preprocessing of sentences and perform analysis to produce results.

4.3.3 Performing Stage

The context-managing technology was developed to extract the requirements of technical specifications. In the Keyword Extraction phase, Parameter, Attribute, Range, and Unit were extracted as regular expressions of Python. In the Context Update (Contextual Renewal) phase, analysis time was reduced because only the SDP data associated with the sentence was optionally performed. Sentences containing keywords from SDP are represented as context scores, depending on their influence on the paragraph. In context score calculation, Parameters 1, 2 and 3 were learned as context keywords and applied to the calculation of context scores in paragraphs. Contextual analysis algorithms are designed to extend the influence of Parameters beyond one sentence to several subsequent sentences. Thus, even if not all parameters in the sentence are included, the context can be grasped and comparative analysis can be performed with SDP.

For example, if a keyword in parameter 1 appears in a sentence, the effect of that sentence is at its highest point. The score decreases as you move on to the next sentence. The sentence is analyzed with relevant contextual information until it eventually becomes zero. The next task is to match the Attribute-Range-Unit Set (ARU Set) to the extracted SDP data. Through the above processes, the results are derived by comparing the embedded SDP data with the extracted SDP data. The results will have True or False if they meet the criteria range, depending on whether they are consistent. The above series of processes are called Context-Managing and a detailed analysis of the results is described in Chapter 6 Analysis of Results.

5 System Integration

It implemented the previously described the algorithm as a system. The configuration of the system consists of three (3) stages, as shown in Figure 3. As mentioned earlier, pre-processed data is required to use the platform's analysis module. Stage 1 utilizes the pre-processing module

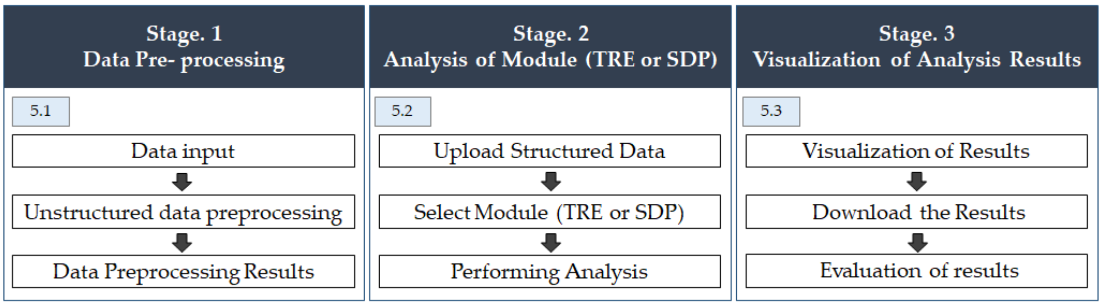

Figure 3. The configuration of the system consists of three stages

of the platform to convert unstructured data (PDF, etc.) to be analyzed into structured data. Stage 2, a full-fledged algorithm analysis is performed. When a user uploads data and selects (directs) an analysis, it refers to a series of processes in which the actual analysis is performed. Stage 3 allows the utilization of the results data as a step to visually present the analysis results on the platform.

Technical specification analysis modules are implemented on cloud service platforms. This cloud service platform is part of the "Artificial Intelligence and Big-data (AI-BD) Platform Development for Engineering Decision-support Systems" project called "Technical Specification Analysis". The platform is called Engineering Machine-learning Automation Platform (EMAP). EMAP allows the selection of classification, regression, and deep learning algorithms for supervised learning, and clustering algorithms for unsupervised learning, and is expected to be completed in the second half of 2021.

5.1 Data Pre- processing

It is the process of selecting technical specifications for analysis and entering them on the platform. Pre-processing the original data through the platform's pre-processing module.

5.2 Analysis of Module (TRE or SDP)

If the structured data is prepared, it is applied to the analysis module (TRE or SDP). First, the corresponding structured data is uploaded to the platform. The next step is to select from the TRE or SDP module of this module and perform the analysis through the algorithm. When the analysis is performed, the screen shows "Performing an analysis" on the screen. At the end of this process, the data frame of the analysis result is constructed and the results are expressed on the screen below.

5.3 Visualization of Analysis Results

The screen expressed by visualizing the analysis results is expressed. Figure 4 is a visualization screen of TRE module analysis results. This histogram shows the severity of the risk in the sentence. Project evaluation information can be obtained from the results summary table on the right. In particular, the evaluation score on the summary table of results is key information in that table. The value is divided into 'total risk score' and 'total number of extracted sentences'.

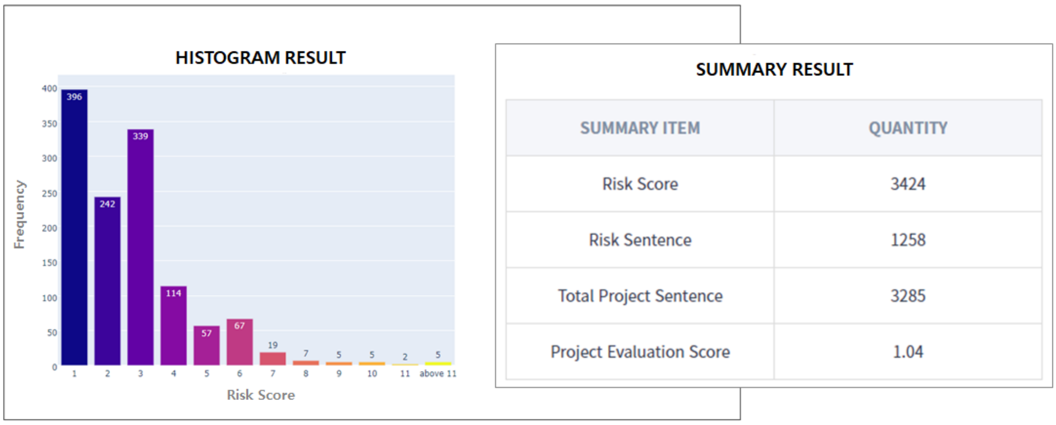

Figure 4. Visualization screen of TRE module analysis results

6 Analysis of Results

The output of the algorithm analysis results can be downloaded and used. The analysis results are checked and evaluated based on the downloaded result file.

6.1 Results of TRE Module

The severity of the risk sentence is shown as a ranking result based on the total score. The result table can check the risk sentence, the position number of the sentence, and the frequency of the

TRL keywords. Through this module, user can know histogram, evaluation score, and summary results.

6.2 Results of SDP Module

The output of the SDP module is as shown in Figure 5. The comparison results are presented as TRUE or FALSE to the right of the parameters of the equipment. You can check the sentences that contain the corresponding numerical requirements.

Figure 5. Analysis results of the SDP module

7 Validation & Conclusion

7.1 Validation

This chapter evaluates extraction accuracy for the Technical Specification Related Development Model (TRE & SDP). Module validation was conducted to ensure availability in EPC projects. Evaluation of information extraction (IE) results via natural language processing (NLP) was performed based on whether relevant information was extracted or irrelevant information (Lee 2018). Pilot test was conducted based on technical specifications of the one (1) representative EPC project. Our research team trained fifteen (15) technical specifications Data and selected one (1) test technical specification separately.

First, we extracted the Risk Extraction value through the system module. Next, the Risk Extraction value was extracted by three (3) engineers currently working on an EPC project. The results of risk extractions were evaluated through five (5) SMEs. Regarding whether the extracted sentence is a true risk clause, the developed module and the engineer each verified. It was conducted based on the TRL and SDP data described in Chapter 4 to assess the reliability of the Risk Extraction values. The data was based on fifteen (15) project technical specifications previously collected based on the criteria established through the evaluation of experts. The total number of sentences for 15 projects is 23,430, of which 7,444 are classified as risk sentences. Based on this, we tested one technical specification data on a module. The pilot test was conducted with Blind Experiment to exclude subjective intervention from engineers.

Table 4. Risk extraction accuracy results of TRE module

		Risk Extraction (Q'ty)	**Extraction Validation (by SME)**	**Extraction Rate (%)**
TRE	System Module	342	314	92%
	Engineer	311	264	85%

As shown in table 4 above, the analysis results of machine learning-based TRE module (92%) show relatively higher risk extraction accuracy than the average (85%) of engineers' execution results.

Table 5. Risk extraction accuracy results of SDP module

		Risk Extraction (Q'ty)	Extraction Validation (by SME)	Extraction Rate (%)
SDP	System Module	187	168	90%
	Engineer	172	151	88%

The performance of the SDP module as shown in Table 5 above was conducted on the parameters of the ten (10) equipment described in Chapter 4. For verification of the SDP module, we defined sentences that meet the Attribute-Range condition as sentences that require risk extraction sentence review. In other words, 'if both attribute and range match' and 'if range is different but comparison range is satisfied within attribute' were calculated as risk extraction values. The SDP module shows that the SDP module is relatively high (90%) compared to the engineer's analysis result (88%). Also Engineers need an average of two (2) to three (3) days to analyze technical specifications, but when using the system module, it can be seen that analysis can be performed within an hour, even including document pre-processing.

7.2 Conclusion

In this study, we developed an algorithm model for technical specifications that requires prior inspection when bidding or performing EPC projects. Two concepts based on machine learning have been proposed to present techniques for verifying the presence of risk and managing project risk severity. The first, TRE module identified the ability to detect and analyze technical requirements or toxin provisions that are prone to omissions or errors in bids due to time constraints or lack of personal competence. The second SDP comparison module compared the numerical requirements of the technical specification standard and the technical specification to be analyzed to confirm that the difference or over-range client requirements can be detected.

To enhance and verify the reliability of the developed module's performance, we are collaborated with EPC project experts from the beginning of the development. The results of a technical specification review by an expert and a contract review by a system module were compared. SMEs from the advisory team who participated in the verification had about 15 to 20 years of experience in carrying out EPC project. In the case of verification, independent verification could be achieved by excluding mutual discussion of the analysis contents. Pilot Test results show that the TRE module-based risk sentence extraction accuracy results are 92% and the SDP module-based risk sentence extraction accuracy results are 90%. Furthermore, we have identified a significant reduction in time spent.

Therefore, the automatic extraction module of technical specifications proposed in this study has the following advantages. First, a technical system was established to more effectively support the task of reviewing technical specifications that are prone to deviations depending on the capabilities of individual managers. Second, we developed a model that enables preemptive risk management by automatically extracting key risks and presenting metrics for evaluation by incorporating machine learning technologies into technology specification analysis algorithms. Finally, the automatic extraction module detects the risk and omission provisions contained in the technical specification, and can be used as evidence for contract consultation with the company and project execution.

7.3 Limitation for Future Research

This study has accumulated data based on the technical specifications for the EPC project in the data collection phase. However, the format or detailed requirements of the technical specifications vary by country and by company. There was a limit to the lack of universal availability when analyzing other types of technical specifications. These problems are common in rule-based information extraction models. To overcome this, the keywords of TRL and SDP must be continuously expanded.

Also, continuous updates are required by building data on various types of technical specifications through machine learning. It is necessary to supplement the technology to extract the requirements present in the table or picture. In addition, it is necessary to build data that efficiently extracts risks from vast technical specifications. If the data accumulated through machine learning is mutually organically applicable to each other, best practices in various fields can be utilized. It is expected that this system can be applied to collaboration as a means of preventing project risk.

8 References

Acknowledgements

The authors acknowledge that this research was sponsored by the Korea Ministry of Trade Industry and Energy (MOTIE) and the Korea Evaluation Institute of Industrial Technology (KEIT) through the Technology Innovation Program funding for "Artificial Intelligence and Big-data (AI-BD) Platform Development for Engineering Decision-support Systems" project (grant number = 20002806).

References

Son, B.Y., Lee, E.B. (2019). Using Text Mining to Estimate Schedule Delay Risk of 13 Offshore Oil and Gas EPC Case Studies During the Bidding Process. *Energies*. 12. 1956. DOI. https://doi.org/10.3390/en12101956

Lee, D.H., Yoon, G.H. & Kim, J.J. (2019). Development of ITB Risk Mgt. Model Based on AI in Bidding Phase for Oversea EPC Projects. *The Journal of the Institute of Internet*. 31:19(4). pp. 151–60. DOI. https://doi.org/10.7236/JIIBC.2019.19.4.151

Sackey, S., Kim, B.S. (2018). Development of an Expert System Tool for the Selection of Procurement System in Large-Scale Construction Projects (ESCONPROCS). *KSCE J Civ Eng*. 22. pp. 4205–4214. DOI. https://doi.org/10.1007/s12205-018-0439-2

Putra, G., Triyono, R. (2015). Neural Network Method for Instrumentation and Control Cost Estimation of the EPC Companies Bidding Proposal. *Procedia Manufacturing*. 4. pp. 98-106, DOI. https://doi.org/10.1016/j.promfg.2015.11.019

Micheli, G., Cagno, E. & Giulio, A.D. (2009). Reducing the total cost of supply through risk-efficiency-based supplier selection in the EPC industry. *Journal of Purchasing and Supply Management*. Volume 15. Issue 3. pp 166-177. DOI. https://doi.org/10.1016/j.pursup.2009.05.001

Zhuang, Y. (2021). Simulation of economic cost prediction of offshore oil pollution based on cloud computing embedded platform. *Microprocessors and Microsystems*. 83. DOI. https://doi.org/10.1016/j.micpro.2021.103993

Fantoni, G., Coli, E. et al. (2021). Text mining tool for translating terms of contract into technical specifications: Development and application in the railway sector. *Computers in Industry*. 124. DOI. https://doi.org/10.1016/j.compind.2020.103357

Ferrucci, D., Brown, E. et al. (2010). Building Watson: An Overview of the Deep QA Project. *AIMag*. 31(3). 59-7. DOI. https://doi.org/10.1609/aimag.v31i3.2303

Saint, D.P. (2018). Mining incoherent requirements in technical specifications: Analysis and implementation. *Data & Knowledge Engineering*. 117. Pp. 290-306. DOI. https://doi.org/10.1016/j.datak.2018.05.006

Oevermann, J. (2018). Semantic PDF Segmentation for Legacy Documents in Technical Documentation. *Procedia Computer Science*. 137. pp. 55-65, DOI. https://doi.org/10.1016/j.procs.2018.09.006

Lee, J.H. (2018) Development of construction risk extraction model for overseas construction projects based on natural language processing(NLP), *Ewha Womans University*. DOI. http://dcollection.ewha.ac.kr/common/orgView/000000148553

The American Society of Mechanical Engineers. Available online: https://www.asme.org/ (15th, June, 2021)

EMAP: Engineering Machine-learning Automation Platform. Available online: http://54.180.148.175/#/intro (12th, June, 2021)