A Licensing Question-and-Answer Tracking Database Using Unsupervised Keyword Extraction

Hyeongseok Eun [a,b*], Dong Hee Kim [a*], Yoonhee Lee [a], Eunkyoung Jee [b], Youngmi Kwon [c]
[a]*KEPCO Engineering & Construction Company Inc., 150-1 Deokjin-Dong, Yuseong-Gu, Daejeon, Korea, 34057*
[b]*School of Computing, Korea Advanced Institute of Science and Technology (KAIST)*
[c]*Dept. Of Radio and Info. Comm. Eng., Chungnam National University, Korea*
[*]*Co-first author: hseun@kaist.ac.kr, doris@kepco-enc.com*
[*]*Corresponding author: ymkwon@cnu.ac.kr*

1. Introduction

Licensing question-and-answer (LQA) entries from regulatory agencies on the safety-related design of nuclear power plants (NPPs) include cumulative data from the time of commencing the design of a Korean Standard Nuclear Power Plant at Saeul NPP Units 3 and 4, which is currently under construction. It contains design data for construction and operation permits and design change data of an NPP during operation. By accumulating and managing LQAs, it is possible to maintain design consistency and safety.

Currently, LQAs are managed simply as a file and rely on users' search ability. However, if an LQA can be indexed and linked, traceability among LQAs can be secured. Using indexed information, various statistics can be calculated, based on which statistics related to the importance of systems and subsystems can also be calculated.

In this study, the main keywords were extracted for LQA using KR-WordRank, which is based on unsupervised learning. LQAs were connected using the extracted keywords, and it was shown that the classified LQAs could be tracked and navigated. In addition, a new software quality metric called "Software Importance" was introduced by calculating statistics using indexed keywords.

2. Background

2.1 Problems

2.1.1 Digitalization problem

The primary problem in creating an LQA database is noncomputerized data. Some LQAs are lost and only maintained as hard copies or PDFs. Hence, it is necessary to computerize LQAs into a readable file.

2.1.2. Lack of traceability and visibility

LQAs are not grouped by topic; therefore, traceability has not been established. This traceability is not a requirement for regulatory agencies. When the licensee or designer needs to respond to licensing inquiries, manual searching of old inquiry information for each NPP is needed.

In general, the user can create this traceability by preparing tables for each question and answer. However, when linking traceability in tabular form, navigation to access the original text as well as calculate visible and statistical information is difficult.

2.1.3 Security issue

Most natural-language-processing solutions using generative artificial intelligence and supervised learning, recently represented by ChatGPT, are provided online. Accordingly, the LQA database has to be connected to the Internet network for using these solutions. However, connecting to an Internet network is difficult because of a security issue that the answer content sometimes contains design information. Therefore, it is difficult to practically use these online solutions in industrial areas such as NPP engineering.

2.1.4 Absence of information on the importance of systems

There is a qualitative difference in importance among safety-critical systems, important-to-safety systems, and non-safety systems and their software. For example, the importance of a plant protection system (PPS) and diverse protection system (DPS) can be inferred from the relationship between primary and backup. However, the importance of an engineered safety features-component control system (ESF-CCS), engineered safety feature actuation system (ESFAS), and a core protection calculation system (CPCS) can vary depending on user perspective. Moreover, bistable processor (BP) programs are considered more complex and important than coincidence processor (CP) programs; however, because these contents are subjective and relative, no attempt has been made to quantify them.

2.2 WordRank

Chen et al.[1] proposed an unsupervised learning method for word recognition called WordRank, which uses mutual reinforcement and provides an extractive summarization algorithm for a document. WordRank creates a much shorter text that covers all the main points of the document without duplication. It solves the optimization problem of selecting sentences using sentence scoring and topic diversity. Unlike supervised

learning method, this unsupervised learning method takes less time, and words can be extracted without building training data. However, WordRank shows poor word-extraction performance in Korean due to different language structures.

2.3 KR-WordRank

Kim at el.[2] proposed a customized WordRank algorithm for Korean, named KR-WordRank, by considering its linguistic characteristics and improving robustness to noise in text documents. KR-WordRank can be used in big-data processing with a keyword extraction function for documents written in Korean.

2.4 SILKROAD

SILKROAD[3] is a commercial application lifecycle management solution that manages, controls, and reports artifacts and work products on an entire development lifecycle. It allows a user to manage requirements including images, tables, and equations.

3. LQATD

3.1 Overall approach

In this study, we develop a user-friendly database system called "Licensing Question-and-Answer Tracking Database" (LQATD). LQATD allows inspectors and designers to track all LQA entries with traceability. It uses KR-WordRank for keyword extraction and can navigate all question-and-answer entries using keywords. We used SILKROAD to prepare traceable QA information for navigation with keywords and titles. LQATD construction involves four main steps:

1) Data computerization: Two thousand three hundred and thirty LQAs, which are available at KEPCO E&C, were computerized. The computerized LQA dataset was limited to inquiries related to the Instrument and Control (I&C) department or some Safety Analysis department. Lost or in-progress LQAs were excluded. The numbers of LQAs for each NPP are given in Table I.

Table I: Computerized LQAs

NPP projects	LQAs
Hanul NPP units 5,6	906
Hanbit NPP units 5,6	130
Shin Kori NPP units 1,2	386
Shin Wolseong NPP units 1,2	67
Sauel NPP units 1,2	690
Sauel NPP units 1,2 design change project*	29
Shin Hanul NPP units 1,2	765
Sauel NPP units 3,4	264
Total	3237

* Note: The project is titled "Project for Prevention of Reactor Shutdown When 12-Finger CEA Drops" for Sauel NPP units 1 and 2.

2) Preprocessing and keyword extraction: Keywords were extracted for each LQA using KR-WordRank. Stopword sets, which are meaningless word sets, were used to remove meaningless information from the result. A single word was divided into syllable tokens for semantic analysis, and the token strings were compared with stopword sets and removed if they belonged to stopword sets.

In addition, in cases where the names of systems were different for each NPP, they were unified into one representative name for statistical calculation. For example, the "digital plant protection system (DPPS)" token was changed to PPS, and the "reactor core protection system (RCOPS)" token was changed to CPCS to unify the terms.

Two tokens for one question were selected by semantic unsupervised learning with KR-WordRank. In addition, the two most frequent words for one question were also selected, resulting in a total of four keywords. Two tokens and two most frequent words for one answer were also selected in the same way, therefore, eight keywords per LQA could be selected.

The unsupervised learning was performed in an offline security network, and the keyword extraction was completed within 1 minute for each NPP.

3) Post-processing: Eight keywords had duplicates; therefore, these duplicates were removed. Additionally, if the keyword was one letter or a meaningless word, it was removed.

4) Traceability connection and document registration: Post-processed keywords and LQAs were registered in SILKROAD as tables and documents. Table II presents a table example with the regulatory question number, title, question, answer, and extracted keywords.

Table II: LQAs with keywords

Question Number	Question Title*	Question Keywords	Answer Keywords
SWN12_1 :IV-I&C-1	Trip Setpoint Change Procedure*	CPCS	Addressable
		CEAC	Constants
		Modification*	PLUS7
		Constants*	Modification*
SKN34_5 4:PSAR-I-7.2-29	Diversity	Diversity*	Diversity*
		SAR	Trip*
		Design*	Channel*
		Testing*	System*

* Note : as Korean

LQAs with keywords were registered in the SILKROAD database. The primary key was the question number, and it was linked to a maximum of 10 properties, including the question title and keywords. Accordingly, the LQAs were connected via a many-to-many connection. LQATD is the database system with keyword traceability as shown in Figure 1.

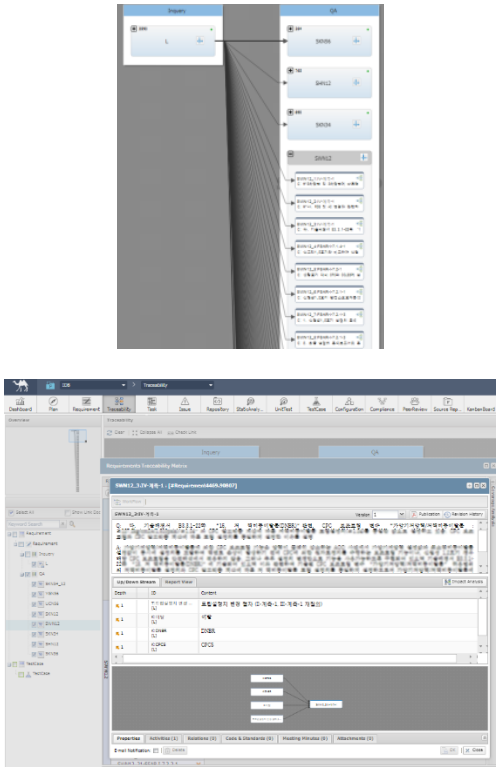
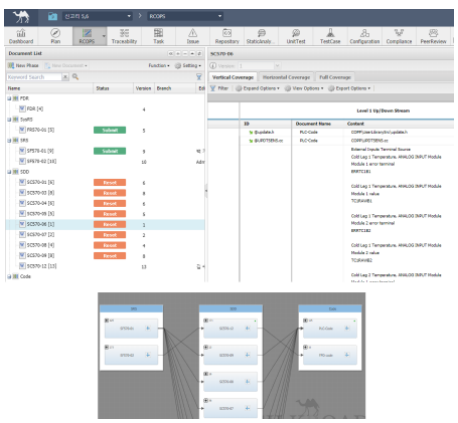


Figure 1. LQATD

If users want to find BP LQAs, they can check for them immediately by tracking sub-questions in the BP keyword; then, they can find the related LQAs in all NPPs. Additionally, by combining with the existing requirement traceability matrix (RTM) for software documents, software implementation details can be checked under enhanced configuration control management as shown in Figure 2.



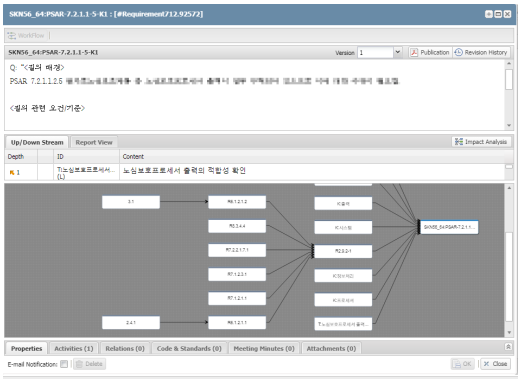


Figure 2. LQATD with RTM

For example, questions about BP can be compared with the LQAs for Shin Wolseong Units 1 and 2 and the LQA for Saeul Units 3 and 4; thus, the response and design consistencies can be checked. Moreover, the software document contents can be tracked to confirm the details of the design.

3.2 Statistical analysis

Because LQATD includes many inquiries from the NPPs, navigations on various keywords are possible. Table III lists the top 20 keywords in LQATD.

Table III: Top 20 Keywords in LQATD

Rank	Keywords	No of LQAs
1	System*	1005
2	PPS	609
3	ESF-CCS	330
4	Test*	328
5	Design*	323
6	Signal*	225
7	Trip*	224
8	Channel*	223
9	Safety*	218
10	Software	213
11	CPCS	171
12	Control*	167
13	Setpoint	152
14	Processor*	148
15	ESFAS	138
16	Data*	126
17	ITP	119
	Digital*	119
19	Logic*	114
20	Communication*	113

* Note : as Korean

As calculating statistical figures for the inquiries and keywords was possible, we focused on the relative ratio of inquiries between the I&C systems; the calculated

figures are given in Table IV for I&C systems, and Table V for subsystems of I&C systems.

Table IV: LQAs for I&C systems

System	No of LQAs
PPS (DPPS, RPS*[1])	609
ESF-CCS	330
CPCS (RCOPS)	171
ESFAS*[1] (DESFAS)	138
DPS	69
COLSS	51*[2]
QIAS-P (ICCMS)	44
QIAS-N	16
SPADES (CFMS)	12
Total	1440

* Note 1 : It is usually used as a function name, but it was unified into a system name for statistical calculation.
* Note 2 : It is a calibrated number.

Table V: LQAs for subsystems of I&C systems

Subsystem	No of LQAs
BP (for PPS, DPPS, RPS)	53
CP*[1] (for PPS, DPPS, RPS)	48
COPP*[2] (for RCOPS)	16
CEAP*[2] (for RCOPS)	5
CCP*[2] (for RCOPS)	6
ITP*[3]	119
MTP*[3]	59

* Note 1 : Because DPPS uses the term "LCL" instead of "CP", the terms are unified into “CP”.
* Note 2 : For the Shin Hanul NPP units 1&2 and Sauel NPP units 3&4
* Note 3 : In the Shin Hanul NPP units 1&2 and Sauel NPP units 3&4, MTP and ITP exist in each system.

The regulatory agency often inquires about the system, not its subsystem or software. Therefore, the sum of the numbers of LQAs for the subsystems is smaller than the number of LQAs for the system.

In the case of core operating limit supervisory system (COLSS), KEPCO E&C typically does not have the primary responsibility for answering COLSS inquiries; thus, it has fewer inquiries than other systems. Therefore, the number of COLSS inquiries is not objective when using all LQAs as a population. Hence, in this case, the LQA ratio in the Sauel NPP units 1,2 design change project mentioned in Table I is used to determine the objective LQA ratio. Because KEPCO E&C has a primary responsibility in this project, we can obtain the objective LQA ratio. The LQA ratio of the COLSS of the CPCS for this project was calculated to be 1:3 (5 and 15 cases, respectively); therefore, the number of LQAs for COLSS was calibrated to 51, which was around 1/3 of the CPCS.

CPCS has CPC, CEAC, and CPP programs. Because the word “CPC” is similar to “CPCS,” it is difficult to distinguish between such keywords. Moreover, an inspector may sometimes confuse these words. For example, if the inspector asks a question using the term “CPC” instead of “CPCS,” it is not a CPC program inquiry, but the “CPC” keyword is extracted by KR-WordRank. Then, it will be incorrect data. CPC, CEAC, and CPP can correspond to COPP, CEAP, and CCP in RCOPS, respectively. Therefore, instead of these words, COPP, CEAP, and CCP, which are the programs in RCOPS, are used to calculate the objective LQA ratio.

3.3 Software Importance index

PPS received the most questions among the pieces of software listed above, at 43.3%, approximately 1.8 and 3.6 times more questions than those for ESF-CCS and CPCS, respectively. The inquiry ratios for BP and CP in PPS were 52.5% and 47.5%, respectively. This result indicates that inspectors questioned BP approximately 1.1 times more than they did CP. Similarly, as the LQA ratios for COPP, CEAP, and CCP were 59.3%, 18.5%, and 22.2%, respectively, the inspectors questioned COPP approximately 3.2 times more than CEAP.

These abovementioned statistics can be used to estimate keyword importance. Software importance classification requires many interpretations and expert analysis. However, any interpretation and expert analysis cannot avoid subjectivity; thus, an external analyzer cannot calculate the same figure. Therefore, these objective statistics can be seen as a new software quality metric for measuring software reliability. In this study, we suggest a software importance index as follows:

- Target Software Importance index

$$= \frac{Number\ of\ Target\ Software\ Queries}{Total\ Number\ of\ Software\ Queries} \times 100\%$$

The calculated software importance indexes for I&C systems are presented in Table VI.

Table VI: Software Importance indexes for I&C systems

Software	Software Importance index
PPS (DPPS, RPS)	42.3 (BP: 22.2, CP: 20.1)
ESF-CCS	22.9
CPCS (RCOPS)	11.9 (CPC: 7, CEAC: 2.2, CPP 2.6 / COPP: 7, CEAP: 2.2, CCP 2.6)
ESFAS (DESFAS)	9.6
DPS	4.8
COLSS	3.6
QIAS-P (ICCMS)	3.1
QIAS-N	1.1
SPADES (CFMS)	0.8

4. Conclusion

In this study, a systematic database for LQAs was constructed by linking keyword traceability, and keyword extraction was performed for 3237 LQAs using KR-WordRank. It was possible to extract keywords from the LQAs for domestic NPPs to classify LQAs on various topics.

By linking LQAs via keywords, it was possible to visually check the LQAs by keywords, enabling the evaluation of response consistency and design consistency. Moreover, by combining with the existing RTM, the detailed software design can also be checked. Various statistical analyses were possible with keyword-classified information. Using the statistical results, we suggest a new software quality index called “Software Importance.” Because this new quality index excludes subjectivity and is based on objective statistical results, more objective software reliability measurement is possible if this new quality index is used for the quantitative evaluation of software reliability.

ACKNOWLEDGMENTS

This research was partly supported by the Nuclear Safety Research Program through the Korea Foundation Of Nuclear Safety (KoFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No. 2105030) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1I1A1A01072004).

REFERENCES

[1] Chen, S., Xu, Y., and Chang, H., “A simple and effective unsupervised word segmentation approach,” In proceedings of the 25th AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 2011.
[2] H. Kim, S. Cho, and P. Kang, "KR-WordRank : An Unsupervised Korean Word Extraction Method Based on WordRank," Journal of Korean Institute of Industrial Engineers, 2014.
[3] SILKROAD RM. http://www.silkroadalm.com/main.do?menu_item=rm_overview (accessed March. 10, 2024).