Ph.D Thesis Dissertation Presentation
2 April, 2018

Hybrid Failure Diagnosis and Prediction Framework in Large Industry

Dohyeong Kim

Department of Computer Science and Engineering
Kyung Hee University

Advised by
Kyung Hee University : **Prof. Sungyoung Lee, PhD**
University of Tasmania : **Prof. Byeong Ho Kang, PhD**

Ubiquitous Computing Laboratory
Kyung Hee University, Korea

KYUNG HEE UNIVERSITY

UNIVERSITY of TASMANIA AUSTRALIA

Table of Contents

Background: Industrial plant maintenance

- The recent trend of industrial plant maintenance focuses on two main factors, **alarms** and **human expertise**.
- **The alarm system collects the status of different types of facilities from the sensors** in each facility and **announces status of facilities to human experts**.
- Experts describes their failure maintenance experience to the **failure report**, and it can be used as **references about other failure**.

Failure reports

Alarms are used *to detect specific symptom* of the facility

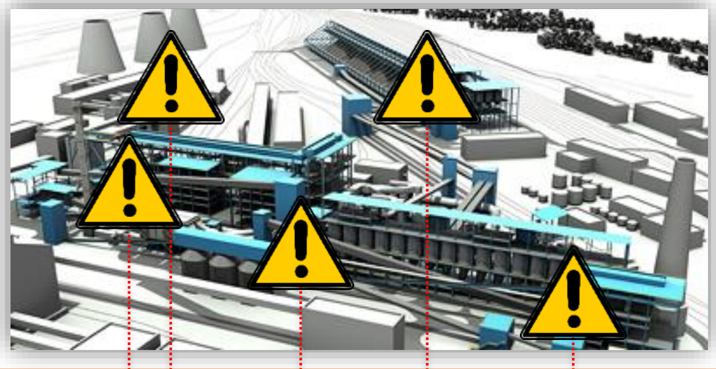
Human experts have *sufficient knowledge in diagnosing and treating failures*

After solving failure, *experts write whole process (cause analysis, treatment action)*

Motivation: Issues in the industrial plant maintenance

There are two issues that should be solved for industrial plant maintenance

- **The system may produce alarm flooding**
  - **Enormous amount of the collected alarm** should be checked and **handled by human experts**.
  - Failures can be misled or skipped → **A critical industrial disaster**
- **Diagnosis and treatment activities are too dependent on human experts**
  - **Only limited numbers of human experts** have sufficient experiences in the certain industrial plant.
  - **Some failures cannot be diagnosed or treated** since the expert have never experienced before [4].
  - **Failure report** aims to use for failure diagnosis and treatment, but in reality it is **difficult to apply for failure management**.



Many and various alarms occur on real-time in plants

Human experts deal with problems by their expertise

Failure reports are difficult to apply for failure management

Problem Statement

> *In order to prevent the huge industrial accident, it is crucial to acquire real-time facility data and analyse the expertise, and computerise them for the intelligent system*
>
> CEO of Tesla, Elon Mask

◆ **Knowledge Acquisition for Failure Detection**

- Machine learning is difficult to acquire clear and proper knowledge to domain and continuous maintenance is not possible.
- Human knowledge engineering is, in initial stage, KB constructing cost is high (slow pace),
  knowledge maintenance cost and the KB size are directly proportional.

◆ **Knowledge Reuse for Failure Diagnosis and Prediction**

- Failure experiences (cause-and-effect of failure) are written in failure reports by experts.
- The reports are written in unformatted manners, but in reality these tend not to use in the failure case maintenance.

Goals

◆ **Discover** the knowledge for failure detection, and **prevent** the failure in the large industrial plants

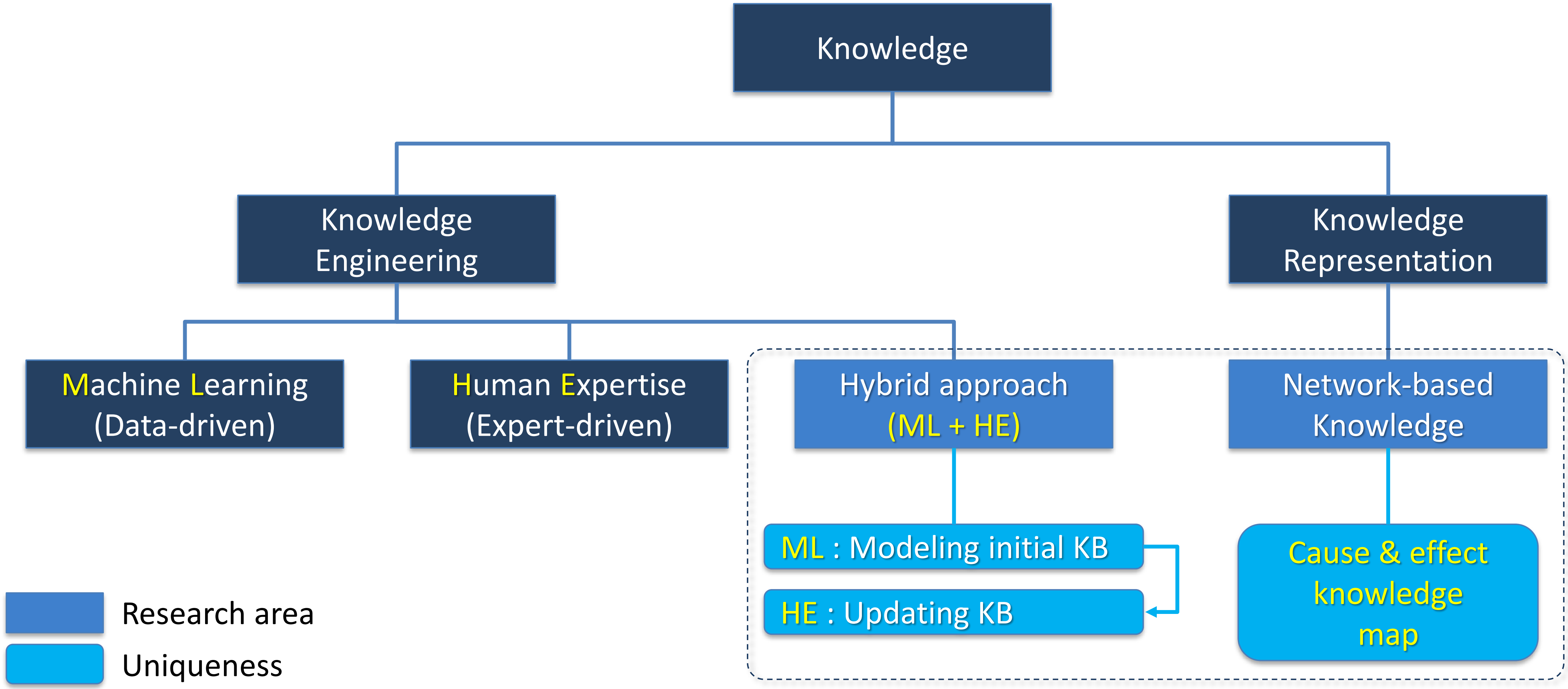
Objectives

◆ To discover the failure detection knowledge by using real-time alarm data and machine learning techniques.
◆ To acquire the failure diagnosis and prediction knowledge from domain expert written failure reports.
◆ To purpose failure Prediction Framework using two knowledge representations

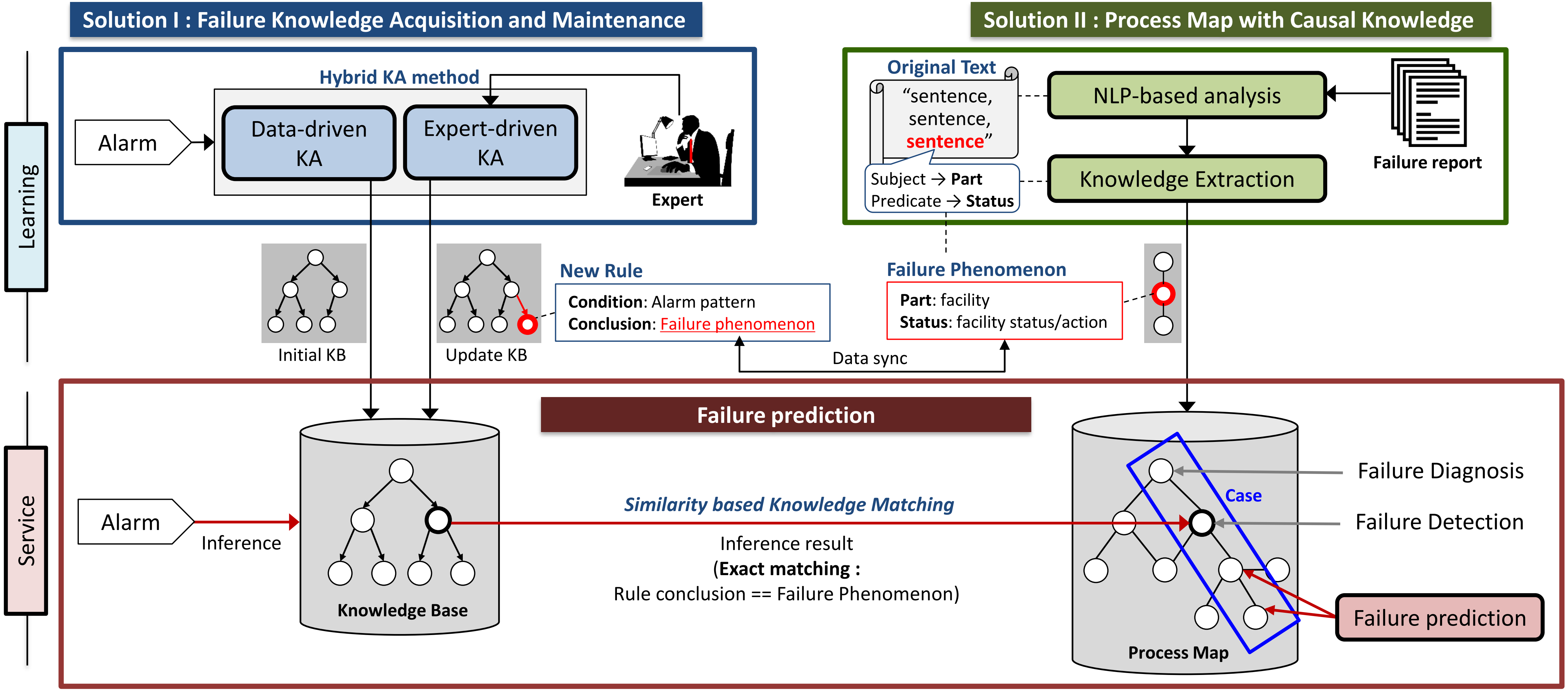
Challenges

◆ Under big data environment, integrating the process of ML knowledge acquisition and human knowledge engineering is crucial.
◆ Acquiring the casual knowledge from the unformatted failure report with unstructured natural language is almost impossible

Research Taxonomy



Overview: Proposed Methodology



UCL Ubiquitous Computing Laboratory
Kyung Hee University, Korea
KYUNG HEE UNIVERSITY
UNIVERSITY of TASMANIA

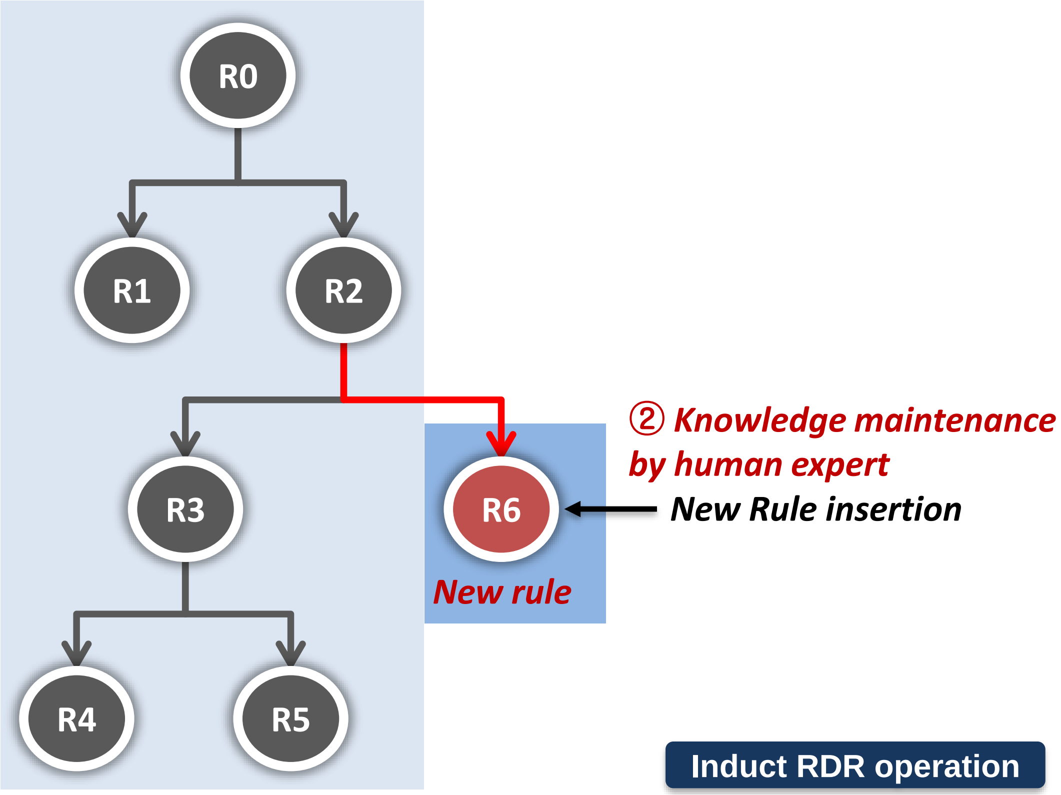
Related work: Failure Knowledge Acquisition and Maintenance

In case of existing machine learning methods, **there are** ***over-generalization and over-fitting issues*** **if the size or range of data is not sufficient**

- **Induct RDR** is a knowledge acquisition approach that ***can be used with human expert*** ***and machine learning*** [1]
- What is **RDR?**
  - ✓ RDR is originally *a tool for acquiring knowledge from human experts*.
  - ✓ RDR supports the function which *enables acquiring the human expert's knowledge based on the current context and adding those knowledge incrementally*
- Why **Induct RDR?**
  - ✓ Induct RDR is a machine learning-based RDR approach, that *allows creating new expertise through machine learning technique*
  - ✓ Induct RDR creates rule in a RDR framework so *it also allows acquiring knowledge from human experts*.

Limitation of the original InductRDR
Produce severe computational issue if the domain has large size of training dataset
If the size of dataset was too large, it is **difficult to distinguish the importance of the rules**
Impossible to handle numerical variable

① Knowledge Acquisition by machine learning (from data)



Induct RDR operation

[1] Gaines, B. R. (1989, December). "**An Ounce of Knowledge is Worth a Ton of Data: Quantitative studies of the Trade-Off between Expertise and Data Based On Statistically Well-Founded Empirical Induction.",** In ML (pp. 156-159).

Related work: Process Map with Causal Knowledge

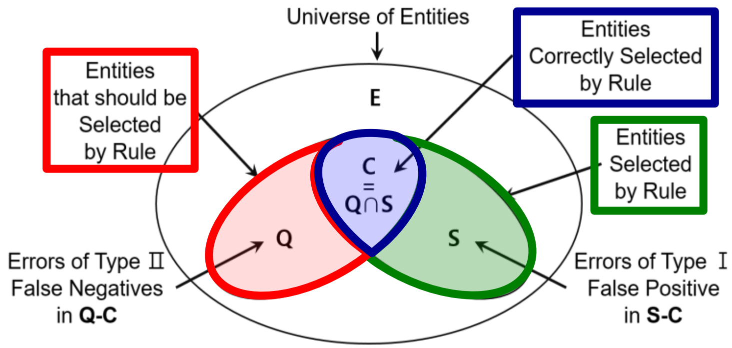
- Proposed Methodology in comparison with ontology engineering tools

	Type of Development	Collaborative Construction	Reusability Support	Degree of Application Dependency	Strategies for Identifying Concepts	Methodology Details	Auto Ontology Building
TOVE [2]	Stage based	X	O	Application semi independent	Middle out	Some Details	X
METHONTOLOGY [3]	Stage based	X	O	Application independent	Middle out	Sufficient Details	X
KBSI IDEF5 [4]	Evolving prototype	X	O	Application independent	Not Clear	Some Details	X
Common KADS and KACTUS [5]	Modular development	X	O	Application independent	Top-down	Insufficient Details	X
ONIONS [6]	Modular development	X	X	Application dependent	Not Clear	Insufficient Details	X
Mikrokosmos [7]	Guidelines	X	X	Application dependent	Rule based	Some Details	X
MENELAS [8]	Guidelines	X	X	Application dependent	Concept Graphs	Insufficient Details	X
SENSUS [9]	Do not mention	O	O	Application semi independent	Bottom up	Some Details	X
Cyc methodology [10]	Evolving prototype	X	O	Application independent	Not Clear	Some Details	X
UPON [11]	Evolving prototype	X	O	Application independent	Middle out	Some Details	X
101 method [12]	Evolving prototype	X	O	Application independent	Developer’s consent	Some Details	X
On-To-Knowledge [13]	Evolving prototype	X	X	Application independent	Middle out	Some Details	X
Proposed method	**Guidelines**	O	O	**Application semi independent**	**Top-down**	**Sufficient Details**	O

Comparisons Original Induct RDR With Updated Induct RDR

	Best Clause Selection	Best Clause Evaluation	Numeric number handling
Core Function	• The original InductRDR searches all possible combinations of terms in order to find the best class.	• The original InductRDR applied m-function, the sum of the standard binomial distribution, for assessing the credibility of the clause	• Use only nominal data
Limitation	• Produce *severe computational issue* if the domain has large size of training dataset	• If the size of dataset was too large, it is almost impossible to use m-values for *distinguishing the importance of the rules.*	• *Can not handle numeric values* • Nominal data can be divided into groups by their values but it is almost impossible to do the same thing for numeric data
Update	• *Sort the terms first* • *Only terms with the smallest m-values* can be *added to the clause*	• *Use Information gain* (key of improving prediction accuracy in decision tree algorithms)	• *Can use numeric values*

m-value for Best Clause Selection



m-value measures **the accuracy of rules**

$$m'(S) = \sum_{i=z}^{\min(s,k)} {}^{s}C_i \left(\frac{k}{n}\right)^i \left(1-\frac{k}{n}\right)^{s-i}$$

n is the number of the whole training set E.
k is the number of the subset Q which contains all the examples which the algorithm needs to learn the rule to select.
s is the number of the subset S which contains all the examples which the rule can actually select.
z is the number of the intersection of Q and S.

When n is too large, the m-value tends to become 0, then all the terms have the same quality.
However, when calculating each attribute, information gains may still show big differences.
Therefore, considering the accuracy of rules in this case,
the best terms must have attributes with larger information gains.

Best Clause Evaluation | Numeric number handling

Numeric data are split into two subsets by calculating information gains.

$$\text{Entropy} = \sum_i -p_i \log_2 p_i$$

Information Gain = entropy(parent) – [average entropy(children)]

p_i is the probability of class i
Compute it as the proportion of class i in the set.

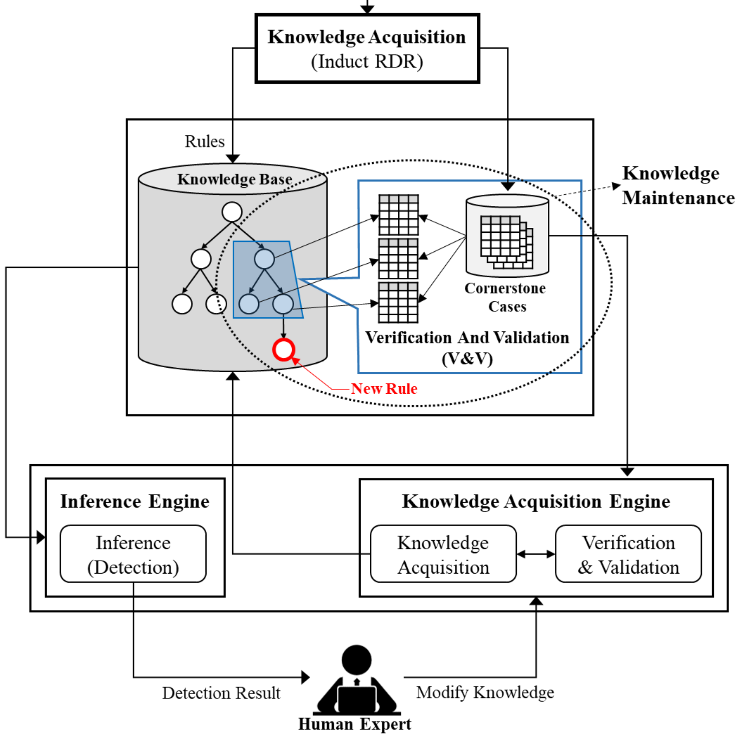
One best rule/clause may contain several numeric and nominal attributes.
The combined clause is still measured by m-value.

[14] **Dohyeong Kim** et al., "RDR-based Knowledge Based System to the Failure Detection in Industrial Cyber Physical Systems", Knowledge-Based Systems (**SCI**, IF 4.529), 2018 (Accepted)

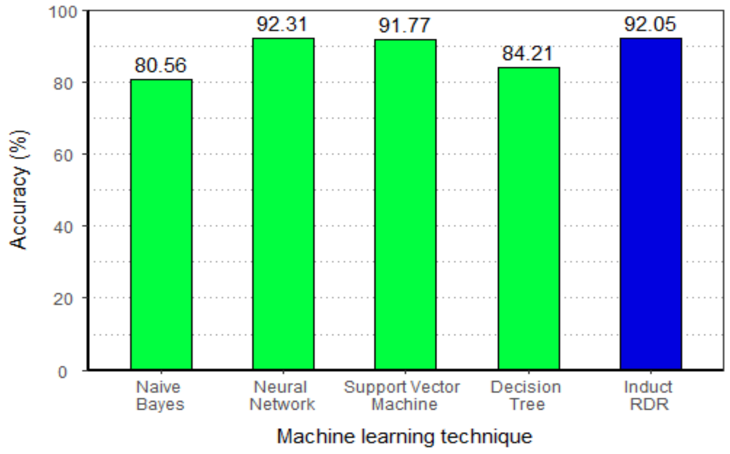
Evaluation : Failure Detection Performance Evaluation

Training Dataset

Attribute						Class
Alarm ID	Time	Facility ID	Count	Lifetime	Ratio	Status
DRV_183	17	H1103364	1	3228	896.67	INTRUDE
ES_041	16	H1101349	10	112	31.11	HUNTING
MCC_323	23	H1103364	1	3600	1000	IMPACT
APC_014	8	H1101349	4	22	6.11	BUR
PAG_004	1	H1101613	13	43	11.94	LEAK
PRC_090	9	H1101349	4	21	5.83	CARBONIZATION
PRC_058	7	H1105709	1	30	8.33	NORMAL



- Data **: 567,748 alarm data** (Hyundai Steel Company)
- Domain experts : **35** (employees in Hyundai Steel Co.)
- **4 algorithms** are selected for comparing with Induct RDR and other algorithms
  - ✓ it shows that Neural Network and Induct RDR achieved over 92% detection accuracy with 10-folds cross validation



The accuracy of failure detection with machine learning techniques

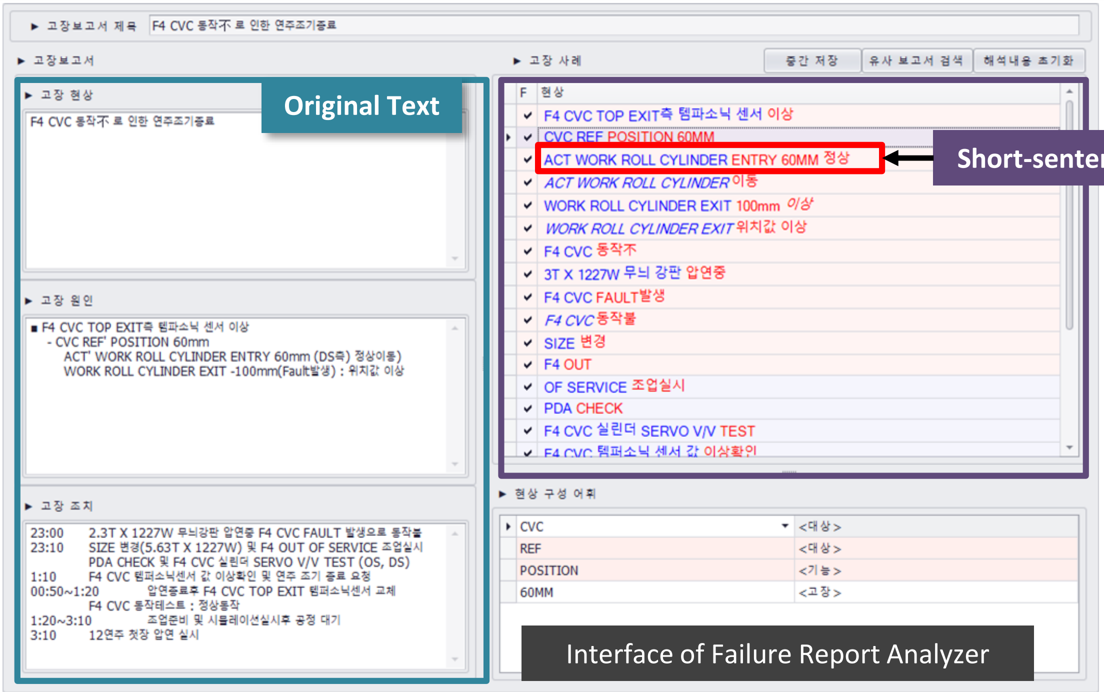
Evaluation Techniques	Detection Accuracy
The updated Induct RDR	92.05%
The updated Induct RDR with human rules	**100%**
Neural Network	92.31%

The performance comparison with machine learning techniques and proposed Induct RDR with human rules

[14] **Dohyeong Kim** et al., "RDR-based Knowledge Based System to the Failure Detection in Industrial Cyber Physical Systems", Knowledge-Based Systems (**SCI**, IF 4.529), 2018 (Accepted)

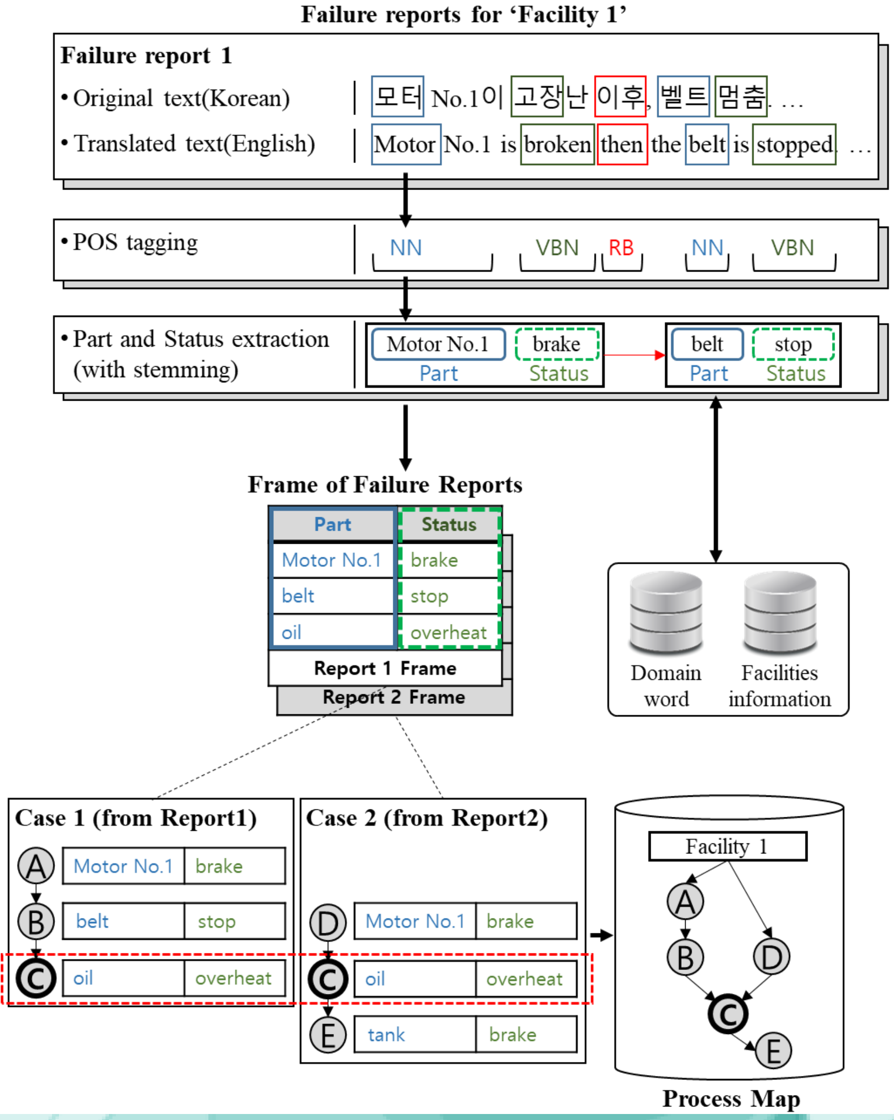
Methodology : Knowledge Extraction Framework

- **Process map**: It represents the relationship among different types of facilities and its failure into the network-based knowledge.
  - ➢ It acquires the cause-and-effect of different facilities' failure and those relationship, and transforms them in a network-based knowledge.
- **Failure report**: When failure occurred from facilities, reasons and treatment action are written by human experts.
  - ➢ A sentence is separated in to **short-sentences** which is consisted of 'Part' and 'Status'.



Part : subject of short-sentence

Status : predicate of short-sentence



[15] **Dohyeong Kim** et al., "A Hybrid Failure Diagnosis and Prediction using Natural Language-based Process Map and Rule-based Expert System", International Journal of Computers, Communications & Control (**SCIE**, IF:1.374), 2018

Ubiquitous Computing Laboratory
Kyung Hee University, Korea
KYUNG HEE UNIVERSITY
UNIVERSITY of TASMANIA

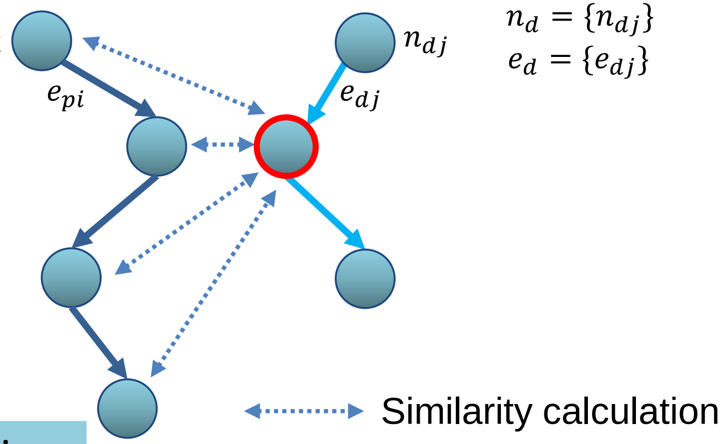
Methodology : Knowledge Modeling

Similarity calculation Between cases

Case 1 $G_p = \{n_p, e_p\}$ Case 2 $G_d = \{n_d, e_d\}$

$n_p = \{n_{pi}\}$
$e_p = \{e_{pi}\}$

$n_d = \{n_{dj}\}$
$e_d = \{e_{dj}\}$



Similarity calculation

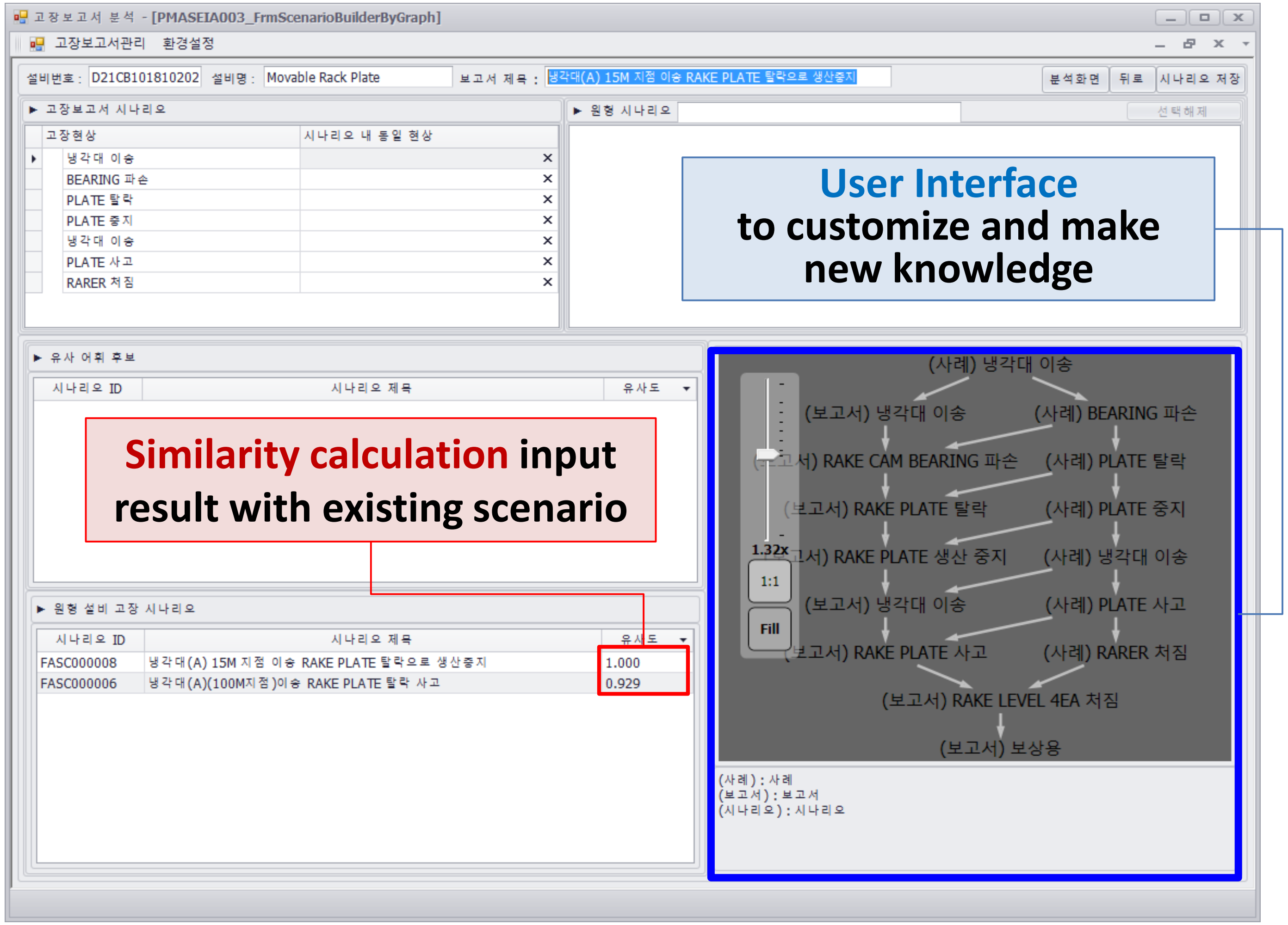
$$Sim(n_{dj}, n_{pi}) = Sim_{Text}(n_{dj}, n_{pi}) + Sim_{Rel}(n_{dj}, n_{pi})$$

Text based similarity

$$Sim_r(n_{dj}, n_{pi}) = EditDistance(n_{djr}, n_{pir}) + EditDistance(n_{djE}, n_{piE})$$

Relation based similarity

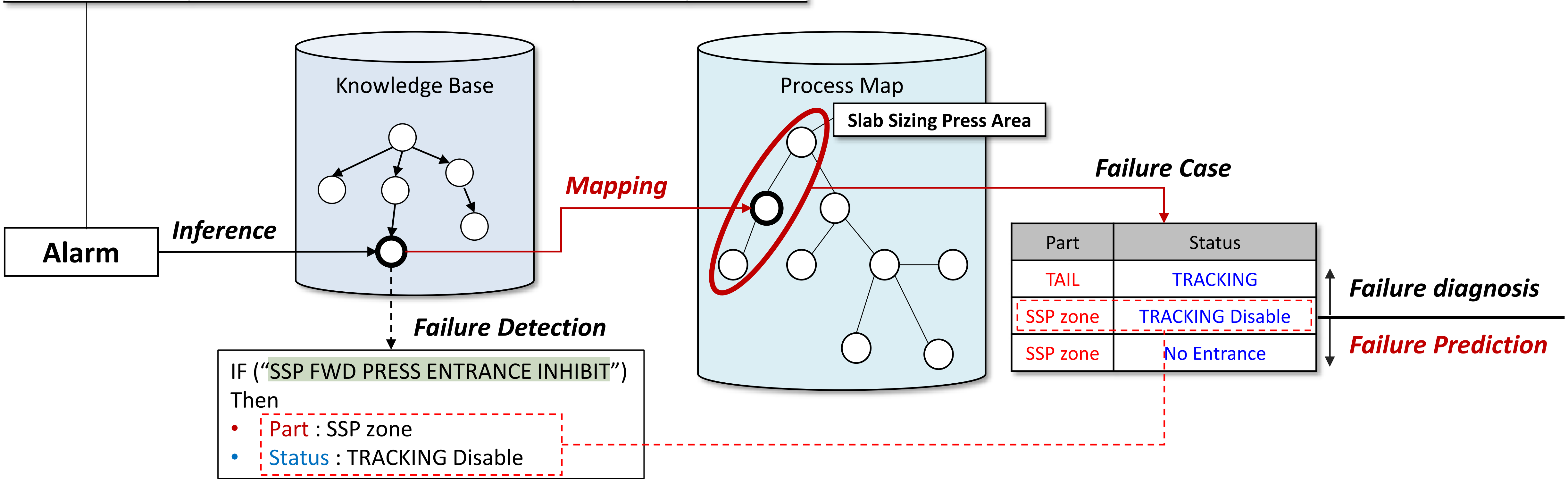
$$Sim_{Rel}(n_{dj}, n_{pi}) = \max_{n_{qk} \in P(n_{qi}), n_{dl} \in P(n_{dj})} Sim(n_{pk}, n_{dl}) + \max_{n_{qk} \in C(n_{qi}), n_{dl} \in C(n_{dj})} Sim(n_{pk}, n_{dl})$$



[15] **Dohyeong Kim** et al., "A Hybrid Failure Diagnosis and Prediction using Natural Language-based Process Map and Rule-based Expert System", International Journal of Computers, Communications & Control (**SCIE**, IF:1.374), 2018

Evaluation : Failure Prediction Framework

Equipment	Alarm message	Count	Lifetime	Ratio
Slab Sizing Press Area	SSP FWD PRESS ENTRANCE INHIBIT	1	3228	896.67
R2 Area	R2 ODD PASS ENTRANCE INHIBIT	10	112	31.11
R2 Area	R2 EVEN PASS ENTRANCE INHIBIT	1	3600	1000
R2 Area	SDD SENSOR SYSTEM UNHEALTHY	4	22	6.11



[15] **Dohyeong Kim** et al., "A Hybrid Failure Diagnosis and Prediction using Natural Language-based Process Map and Rule-based Expert System", International Journal of Computers, Communications & Control (**SCIE**, IF:1.374), 2018

Evaluation : Failure Prediction Performance

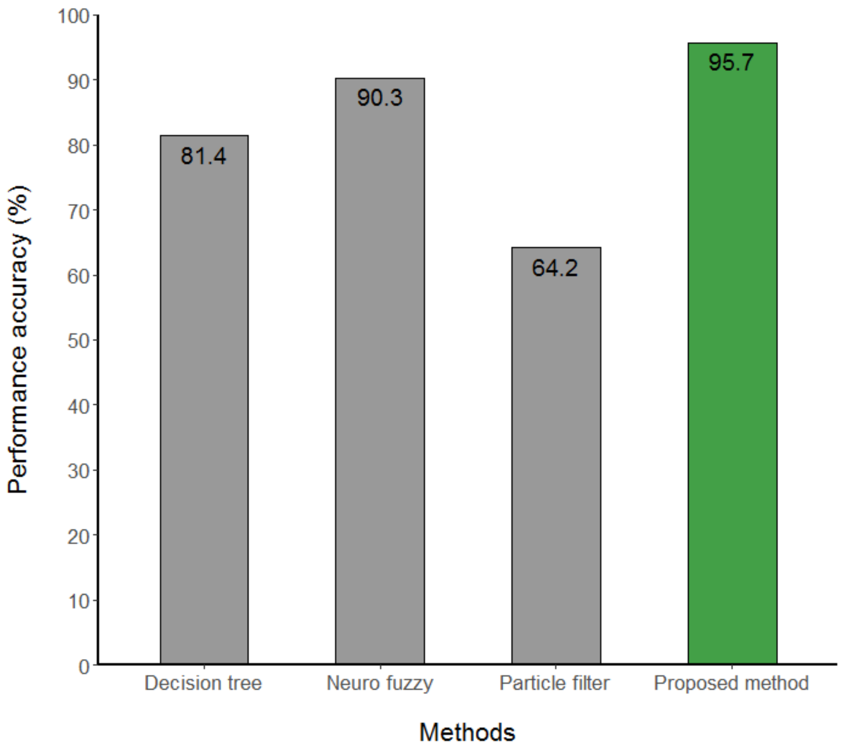
- Data : **400 failure reports, 502,308 alarm data** (Hyundai Steel Company)
- Domain experts : **35** (employees in Hyundai Steel Co.)
- Test data : **100 failure case, 200,923 alarm data**
- Knowledge base : **237 rules**

	Inference	**Failure prediction**
Success Rate	99.1%	98.3%

Success rate of knowledge use

Author	**Description**	**Accuracy**
Santos et al. (2010)	Applied different machine learning techniques (incl. Bayesian Network, SVM, and decision tree)	81.4%
Liu and Jiang (2008)	Used particle filter with Bayesian Inference	64.2%
Chen et al. (2015)	Applied knowledge-based neural fuzzy inference	90.3%
Proposed System	**Natural Language-based Processing Map + knowledge-based alarm prediction system**	**95.7%**

Review of Failure Prediction By Previous Failure Prediction System



Comparison of accuracy of failure prediction

[15] **Dohyeong Kim** et al., "A Hybrid Failure Diagnosis and Prediction using Natural Language-based Process Map and Rule-based Expert System", International Journal of Computers, Communications & Control (**SCIE**, IF:1.374), 2018

Conclusion

This thesis contributes to

[Solution 1]

- proposed a knowledge capturing approach for failure detection that leverages the benefits of machine learning and human experts
  - ✓ Machine learning: reduce time and cost
  - ✓ human experts: minimize the over-generalisation and over-fitting issue
- updated an RDR-based machine learning approach in order to optimize the real-time and big data-based machine learning model by human expertise

[Solution 2]

- proposed a network-based knowledge acquisition approach that enables to acquire and store the network-based knowledge
- the proposed approach allows to update the cause-and-effect network-based knowledge by applying natural-language processing techniques and increment rule acquisition technology

[Proposed Failure Prediction Framework]

- achieved high failure prediction accuracy than other three methods.

Publications

Journal : 10

SCI/E	First author 2 : 1(SCI) / 1(SCIE)	Co-author 5 : 2(SCI) / 3(SCIE)
Non SCI/E	First author 3	

First author
- **SCI** : Elsevier, Knowledge-Based Systems (IF: 4.529, Accepted, 2018)
- **SCIE** : International Journal of Computers, Communications & Control (IF: 1.374, Accepted, 2018)

Conference : 8

International	First author : 2	Co-author : 1
Domestic	First author : 4	Co-author : 1

Total publications : 18

First author : 11

References

[1] Gaines BR. “An Ounce of Knowledge is Worth a Ton of Data: Quantitative studies of the Trade-Off between Expertise and Data Based On Statistically Well-Founded Empirical Induction.”, In ML 1989 Dec 1 ,pp. 156-159

[2] Gruninger, M. and M.S. Fox, 1995. Methodology for the design and evaluation of ontologies. Proceeding of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI.

[3] Fernández, M., 1996. CHEMICALS: Ontology of chemical elements. Final-Year Project. Faculty of Informatics at the University of Madrid.

[4] KBSI, 1994. The IDEF5 ontology description capture method overview. KBSI Report, Texas

[5] Schreiber, G., B. Wielinga and W. Jansweijer, 1995. The KACTUS view on the’O'word. Proceeding of the IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing, pp: 159-168.

[6] Gangemi, A., G. Steve and F. Giacomelli, 1996. ONIONS: An ontological methodology for taxonomic knowledge integration. Proceeding of the Workshop on Ontological Engineering, ECAI-96, Budapest, pp: 95.

[7] Mahesh, K., 1996. Ontology Development for Machine Translation: Ideology and Methodology, Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi = 10.1.1.47.3449.

[8] Bouaud, J., B. Bachimont, J. Charlet and P. Zweigenbaum, 1994. Acquisition and structuring of an ontology within conceptual graphs. Proceedings of Workshop on Knowledge Acquisition using Conceptual Graph Theory, University of Maryland, College Park, MD, 94: 1-25.

[9] Swartout, B., R. Patil, K. Knight and T. Russ, 1996. Toward distributed use of large-scale ontologies. Proceeding of the 20th Workshop on Knowledge Acquisition for Knowledge-Based Systems, pp: 138-148.

[10] Lenat, D.B. and R.V. Guha, 1990. Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project. Addison-Wesley, Addison-Wesley Publishing Company, Inc., Reading, Massachusetts.

[11] Nicola, A.D., M. Missikoff and R. Navigli, 2005. A proposal for a unified process for ontology building: UPON. Proceeding of the Database and Expert Systems Applications, pp: 655-664.

[12] Noy, N. and D. McGuinness, 2001. Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report.

[13] Sure, Y., S. Staab and R. Studer, 2003. On-To-Knowledge Methodology. In: Staab, S. and R. Studer (Eds.), Handbook on Ontologies. Springer, Berlin, pp: 811, ISBN: 3540926739

[14] **D.Kim, Y.Lin , S.Y.Han, B.H.Kang and S.Lee, “RDR-based Knowledge Based System to the Failure Detection in Industrial Cyber Physical Systems”, Knowledge-Based Systems (SCI, IF 4.5), 2018 (Accepted)**

[15] **D. Kim, Y.Lin, B.H.Kang, S.Lee and S.C.Han, “A Hybrid Failure Diagnosis and Prediction using Natural Language-based Process Map and Rule-based Expert System”, International Journal of Computers, Communications & Control (SCIE, IF:1.3), 2018 (Accepted)**

Thank you
for your attention

Q & A ?