A Model-Based Approach to Extract Health Information from Textual Data

D. Mandelli and C. Wang

diego.mandelli@inl.gov, congjian.wang@inl.gov

14th Annual Conference of the Prognostics and Health Management Society
Nashville, November 1st – 4th, 2022
• **Context:** ER data generated by nuclear power plants
  - Examples: monitoring data, condition report, corrective actions
  - Heterogenous formats
    - Textual (events, logs)
    - Numeric (e.g., pump oil temperature)
    - Other (e.g., images)
  - The integral analysis of all data elements provides an accurate representation of asset health and performance

• **Goal:** Assist system engineers to analyze ER data (numeric and textual)

• **This paper:** Extract knowledge from textual data
  - Identify causal relations between events

• **Our work:** Causal reasoning applied to ER data
  - Data is not enough: models are needed
  - Merge two perspectives: System engineer and data scientist

Are both these elements adequately analyzed simultaneously?
ER Data Analytics: Causal Reasoning
ER Data Analytics and MBSE

- Need to emulate system engineer knowledge about components and systems
- **Solution:** Model-Based System Engineering (MBSE) diagram-based representation
  - Identify causal dependencies (links) between “Form” and ”Function” elements
- Link to numeric and textual monitoring data can be easily established
- **MBSE language:** Object Process Methodology (OPM)
  - What about SysML?
- **Workflow**
  - OPM models are created for desired assets and systems
  - Translation into graph structure (i.e., networkX-multiDiGraph)

![Diagram showing relationships between fluid, pressure, temperature, vibration, and system components]
Linking ER Concepts to OPM Models

- OPM diagrams provide a clear link to typical ER concepts
  - Failure modes (function)
  - Ageing and maintenance activities (form)

- Are OPM diagrams sufficient?
  - No: quantification of OPM links is missing
  - Statistical analysis and machine learning are the key to quantify links

- Is modeling & simulation needed?
  - First principle laws (conservation, equations of state)

- Causal reasoning directions
  - Precursor analysis
  - Cause-effect analysis
Causal Reasoning

- **Precursor analysis**: Identify the event that triggers following events
  - OPM models are used to create a graph among ER data elements
  - Available anomaly detection and diagnostic methods can be employed to quantify graph edges

- **Cause-effect analysis**: Identify form element(s) that have caused the precursor
  - OPM models provide information on the form elements that support the precursor (function or attribute)
Analysis of Textual Data

• Information extraction from textual data using Natural Language Processing (NLP) methods
  − Rule based text processing
  − Machine learning for relation identification

• Bounding the analysis
  1. Event report (e.g., anomalous behavior or a corrective action)
  2. Cause-effect relation between events
  3. Temporal relations between events

• Under the hood
  − Open-source NLP libraries (Spacy, NLTK)
  − BERT: transformer-based NLP model (ongoing)
Analysis of Textual Data: NLP Workflow (1/3)

• **Step 1: Sentence segmentation and tokenization**

• **Step 2: Grammar check**
  - Identify typos
  - Identify component/asset IDs (list of allowed IDs might not be available)
  - Identify acronyms and abbreviations: this is the most critical point (40-55% accuracy)

• **Step 3: Part of speech (POS) tagging and lemmatization**

• **Step 4: Named entity recognition (NER):** Database of entities has been created
  - Focus on nuclear power plants
    • Entity types: components, assets, systems, materials, chemical, reactions, …
    • Entity nature: electric/electronic, hydraulic/pneumatic, mechanical, structural, architectural, I&C, …
    • Entity property: measured/observed quantities (with associated unit of measure)
    • Link to OPM entities
  - Sources: Available online databases, domain experts
  - NER testing with actual plant text and NRC reports (>94% accuracy)
Analysis of Textual Data: NLP Workflow (2/3)

• **Step 5: Identification of text attributes** (rule-based NLP pipeline)
  − Location and temporal
    • Time of occurrence, duration
    • Concurrency of events
    • Sequence/order of events
  − Measured quantities
    • Numeric value identification (text, number)
    • Unit identification
  − Testing with actual plant text and NRC reports: 97% accuracy

• **Step 6: Coreference resolution** (Spacy-coreferee)
  − Testing with actual plant text and NRC reports: 77% accuracy

• **Step 7: Conjecture identification** (rule-based NLP pipeline)
  − Events that could have happened in the past
  − Events that might happen in the future

Moerchen, F. 2007. “Unsupervised Pattern Mining from Symbolic Temporal Data.” ACM SIGKDD Explorations Newsletter 9: 41–55.

“Several cracks on pump shaft were observed; they could have caused pump failure within few days.”
• **Step 8: Identification of nature of paragraph** (rule-based NLP pipeline)
  − Starting point: set of nouns, verbs, adverbs, adjectives, relations, transition words
  − Identification of negations, passive forms
  − Health status

| Relation                              | Example                                      |
|---------------------------------------|----------------------------------------------|
| Subj + “status verb”                  | Pump was not functioning                     |
| Subj + “status verb” + “status adjective” | Pump performances were acceptable            |

− Causal relation

| Relation                                                                 | Causal relation       |
|--------------------------------------------------------------------------|-----------------------|
| Event_A + “causal verb” (active) + Event_B                               | Event_A → Event_B     |
| Event_A + “causal verb” (passive) + Event_B                              | Event_B → Event_A     |

− Testing with NRC documents, available plant data, and medical available datasets
  • Health status: 89% accuracy
  • Causal relation: 95% accuracy

• **Step 9: Identification of relations using Bert** (ongoing)
  − Graph structure between entities
  − Search for synonyms/antonyms (Spacy-wordnet)
Analysis of Textual Data

• Example
  – “Bearing failure of CCW Pump 1B caused reduced flow.”

• NLP syntactic analysis
  – Sentence segmentation and word tokenization
  – Part of speech tagging
  – Named entity recognition

• NLP semantic analysis
  – Rely on component and system OPM models
  – Generated causal graph

(CCW Pump 1B, bearing, failed)

(CCW Pump 1B, internal high v flow, degraded)
Final Remarks

Data space

Knowledge space

OPM models

NLP Methods

Precursor & cause-effect analysis

Recorded CCW pump 80% reduced flow

Recorded loss of fluid from RCP-2 seal

RCP-2 seal temperature increasing

RCP-1 seal temperature increasing

Recorded anomaly on CCW pump rpm

CCW pump motor bearing degradation

Recorded CCW pump 80% reduced flow

SSC1 health

Event 2

Event 3

Event 4

SSC2 health

SSC3 health

Recorded loss of fluid from RCP-2 seal

CCW pump bearing, and RCP seal restored