SANAYOJAN: A Framework for Traceability Link Recovery between Use-Cases in Software Requirement Specification and Regulatory Documents

Ritika Jain, Smita Ghaisas and Ashish Sureka

RAISE 2014 Presentation
Presentation Outline

• Research Motivation & Aim
• Related work & Research Contributions
• Research Framework & Experimental Dataset
• Solution Approach
  - Automated Extraction of System Use Cases
  - Automated Traceability Links Recovery
• Performance Evaluation & Results
• Limitations
• Conclusion
Presentation Outline

• Research Motivation & Aim
• Related work & Research Contributions
• Research Framework & Experimental Dataset
• Solution Approach
  - Automated Extraction of System Use Cases
  - Automated Traceability Links Recovery
• Performance Evaluation & Results
• Limitations
• Conclusion
System Use Case

- High level description of system processes
- Actor performs an action to achieve a system’s goal.
- Ensures that the behavior of the system is what the user requires.
Compliance Detection

• Organizational projects need to comply with the certain regulations and laws set up by regulatory authority.

• Failure to comply with legal laws and regulations may lead an organization to face penalties and litigation risks.
Research Motivation

• Automated System Use Case Extraction

- Identifying elements of text for legal compliance from documents written in natural language.
- Have no particular format.
- Embedded between other elements of text like definitions, examples, purpose, scope of document.
- Consumes time and cost if done manually.
Research Motivation (Contd.)

• Traceability Links Recovery
  - Manual process is not scalable.
  - Terminology mismatches.
  - Hard to keep track of changing regulations.
  - Time consuming.
Research Aim

• Investigate use of certain features to automate the extraction of system use cases from URS documents.

• Examine similarity based techniques to find traceability links between system use cases and regulations.

• Conduct empirical study on real-world dataset and validate the proposed approach.
Presentation Outline

✓ Research Motivation & Aim
  • Related work & Research Contributions
  • Research Framework & Experimental Dataset
  • Solution Approach
    - Automated Extraction of System Use Cases
    - Automated Traceability Links Recovery
  • Performance Evaluation & Results
  • Limitations
  • Conclusion
# Related Work

<table>
<thead>
<tr>
<th>Researcher</th>
<th>Research Aim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breaux et al.</td>
<td><strong>Semantic parameterization</strong> to restate regulations in RNLS. Identify rights and obligations conflict.</td>
</tr>
<tr>
<td>Huang et al.</td>
<td>Use <strong>probabilistic network model</strong> in computing similarity scores.</td>
</tr>
<tr>
<td>Maxwell et al.</td>
<td>Use <strong>prolog statements</strong> to validate regulations with requirements.</td>
</tr>
<tr>
<td>Massey et al.</td>
<td>Evaluate legal compliance of security requirements using <strong>direct terminology mapping</strong>.</td>
</tr>
<tr>
<td>Xaio at al.</td>
<td>Use <strong>linguistic analysis</strong> to extract access control policies from NL documents.</td>
</tr>
<tr>
<td>Ghaisas et al.</td>
<td>Use <strong>linguistic analysis</strong> to extract rules from SRS documents.</td>
</tr>
</tbody>
</table>
Research Contributions

• Automated extraction of system-use cases from URS documents.

• Application of lexical and semantic based similarity computation algorithms for traceability link recovery between system use-cases and regulations.

• Conduct series of experiments on real-world URS and regulatory documents and present performance evaluation results of the proposed approach.
Presentation Outline

✓ Research Motivation & Aim
✓ Related work & Research Contributions
• Research Framework & Experimental Dataset
• Solution Approach
  - Automated Extraction of System Use Cases
  - Automated Traceability Links Recovery
• Performance Evaluation & Results
• Limitations
• Conclusion
Research Framework
## Experimental Dataset

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num of SRS documents</td>
<td>69</td>
</tr>
<tr>
<td>Average num of pages in the SRS documents</td>
<td>40</td>
</tr>
<tr>
<td>Total num of system use-cases in dataset</td>
<td>1518</td>
</tr>
<tr>
<td>Average num of use cases per document</td>
<td>22</td>
</tr>
<tr>
<td>Num of Regulatory documents</td>
<td>16</td>
</tr>
<tr>
<td>Average num of pages in regulatory documents</td>
<td>50</td>
</tr>
<tr>
<td>Average num of regulations per document</td>
<td>48</td>
</tr>
</tbody>
</table>
Experimental Dataset (Contd.)

- Data Set characterization.
- 10% documents do not contain system use cases.
- Different lengths of documents with varied number of terms.
Presentation Outline

- Research Motivation & Aim
- Related work & Research Contributions
- Research Framework & Experimental Dataset
  - Solution Approach
    - Automated Extraction of System Use Cases
    - Automated Traceability Links Recovery
  - Performance Evaluation & Results
  - Limitations
  - Conclusion
Solution Approach

Phase 1 - Identification of Features

• A system use case is a high level description of system processes where an actor performs and action to achieve system’s goal.

• Manually inspect documents to identify distinguishing features of system vs. non-system use case.
Feature Types

**Spatial**
- Tabular Representation
- Diagrammatic Representation
- Structural representations

**Lexicon-Based**
- Use Case Specific Terms
- Domain Agnostic/Specific Terms

**Syntactic**
- Conditional Use Cases
- Event-Triggered Use Cases

**Miscellaneous**
- Combination of spatial, lexical and syntactic features
Number of columns of table varies from 2-4.
1-2 column contains less than 5 terms of text (the System Use-Case ID and label) and another column containing more than 9 terms of text (the System Use-Case description)
- 98% documents do not contain Use Case Diagrams and the documents containing use case diagrams are followed by textual description of use case as well.

- Ignore Diagrammatic representation in System use case extraction.
Spatial level of text alone cannot distinguish system use case from non-system use cases.
Lexicon-Based Features

“Use Case Specific Terms”

1.3.1 Process Activity-1 Receive notification of death claim

1.3.1.1 Process description

The notification of death may be received along with the death certificate or at times the notification may not be received along with the death certificate. The notification of death (NOD) may also be received from the Independent Financial Advisor (IFA) or in the solicitor information form. Check if the policy number/s is/are quoted in the notification else check for policy number/s by doing a search or call up and obtain the policy number/s. The notification of death may also be received in the form of a letter from the claimant.

Initiate Settlement Process.

The use case starts when the assessment worksheet or the interim payment request is approved. On approval the assessment/interim request work item would be routed to a pool of Claim Officers for Claim settlement.
"Insurance Domain Noun Terms"

**Lexicon-Based Features**

- **USCT**
- **IDNT**
- **DAAT**
- **UIST**

**Noun - Lexicon**
- Member
- Insurer
- Claimant
- Reinsurer
- Settlement- Officer
- Deceased
- Insurance Company
- Handler
- Insured

**Reinsurance:**

If the policy is under Reinsurance cover then the Reinsurer is intimated of the Death claim.

**Process Activity - 7 Verify Previous Part Surrender/withdrawals**

1.3.7.1 Process Description

Withdrawals or previous part surrenders settled earlier, if any, to the deceased needs to be checked in the withdrawals database.
**Domain Agnostic Action Terms**

**Lexicon-Based Features**
- USCT
- IDNT
- DAAT
- UIST

**Action Verb – Lexicon**
- Notify
- Update
- Modify
- Generate
- Indicate
- Approve
- Receive
- Validate
- Display

**Example Usage**

1. **Receive Death Intimation:**
   
   Intimation about the death of the life assured is received by the Insurance Company through different channels like phone call / written communication from the Agent, friend, relative, etc. Once the intimation is received the same should be recorded and the status in the policy and the client master should be updated accordingly to stop processing of generation of premium notices, etc.

2. **Display Reserve Allocation**
   
   2.1 The user indicates to view the reserve allocation
   2.2 The system will display the latest Claim reserve allocated.
Lexicon-Based Features

USCT
IDNT
DAAT
UIST

“User Interface Specific Terms”

Lexicon
Screen
Enter
Display
Select
Choose
Scroll
Edit
Click
.
.

6.3.1.2. Policy – Premium Details
This section displays the plan premium details. For data elements, refer ‘Premium details’ section from "IPP & GPP Policy View" tab of data requirement catalogue (refer section 7).

6.3.1.6. Policy – Income and Withdrawals
The CSR can then click on the details of a particular Income/Withdrawal transaction to view the details.
Conditional Patterns

- If ... Then ...
- Where ... There
- <Activity> only if
- Check if ...
- Validate if...
- 
- 

POS Patterns

- *IN * VB * RB * VB*
- *W1 * VBN * RB *
- *VB * MD* VB *

If [IN] the policy is [VB] in trust then [RB] all the trustees must sign [VB] the death claim application.

Joint life policies are accepted [VB] only [RB] if [IN] both the signatures are available on the form.
Event Triggered Patterns

Before ...
After ...
Once <action>
When ... then ...
In order to ...
In case of ...
...

POS Patterns

* MD * VB * IN *
* WRB*NN*VBN
*
*
*

1.3.8 Process Activity 8: Generate Chargeable Event Certificate

1.3.8.1 Process Description

When a life policy is encashed at surrender, maturity or earlier death of the assured it is common to make profit i.e. when the amount paid by the company exceeds the total amount of premiums paid by the policyholder. Under such circumstances the policyholder is subject to Income Tax on these policy profits. These are known as chargeable gains.

Case must be referred at the early stage of the claim, after [MD] notification [VB] of death and on [IN] receipt of death certificate.
1.3.5 Process Activity: Verify Previous Withdrawals

1.3.5.1 Process Description

The policy could have been entitled for earlier withdrawals and withdrawals could have been processed. Hence a check on the previous withdrawals made is required.

For descriptive texts, check for syntactic patterns
“Use Case Parts Identification”

Miscellaneous Features

- SUCI
- UCPI
- NAUI

Spatial + Syntactic

- at same level
- Level 3
- Normal Text without heading level

**Rule 3:** Claim by Risk Requirements

d) If it is Disability Claim
   
   i) File Policy enquiry Report
   ii) File First Information Report
   iii) File Physician’s Report
   iv) File Hospital records

Combine bulleted texts, check for syntactic patterns
“Noun Action Use Case Identification”

Spatial
Length Check on number of terms >=10

POS Patterns
Remove Stop words
NN * VB

8.1.2 Switch out process
- System sends a communication to the FA with the details of number of units for each fund before the pre-defined cut off time in accordance 2.1/B/COM/CM 007.
- FA completes the Sale transaction based on the communication received.
- FA notifies the SA with the details of the sale transaction.

2. Generate Offer Letter/ Discharge Voucher
   2.1. User indicates Offer letter and Discharge Voucher are not applicable or not
Phase 2 – Automated Extraction using identified features

• Extract **System Use Cases** from URS using features identified.

• All features are used in combination over complete dataset.

• Save them in separate location.

---

**A sample URS document**

---

---

---
Presentation Outline

- Research Motivation & Aim
- Related work & Research Contributions
- Research Framework & Experimental Dataset
- Solution Approach
  - Automated Extraction of System Use Cases
    - Automated Traceability Links Recovery
- Performance Evaluation & Results
- Limitations
- Conclusion
Phase 3 – Individual Regulation Extraction

• Every statement is important to consider.

• Relatively insignificant to automate extraction of regulatory statements.

• Manually write regulations - new line separated statements.

12. Services offered by TPA in relation to Health Insurance Policies

a. The insurer may enter into an Agreement for the provision of defined services with a TPA holding a valid license issued in accordance with the IRDA (Third Party Administrators) Regulations, 2001 as may be amended from time to time.

b. The services offered by a TPA shall not include

   i. Claim settlements and rejections with respect to the health insurance policies; However, TPA may handle claims admissions and recommend to the insurer for the payment of the claim settlement, provided a detailed guideline is prescribed by the insurer to the TPA for claims assessments & admissions in terms of capacity requirements, internal control requirements, claim assessment & admissions procedure requirements etc under the agreement.

   ii. Any services directly to the policyholder or insured or to any other person unless such service is in accordance with the terms and conditions of the Agreement entered into with the insurer.

c. The TPA shall have in place the infrastructure necessary to extend the health services as required to the policyholders at all times.

d. Settlement and Denial of Claims:

   i. All documents submitted to TPA shall be electronically collected and shall be forwarded to the Insurers for taking a decision on the claim settlements or claim rejections.

   ii. TPA shall, in the correspondence to the policyholder with respect to settlement/denial of the claims, state clearly the following:

   "As per the instructions of the insurer <Name of the Insurer>, the claim is being settled/denied for Rs. <amount> on account of <specifies of treatment/grounds of denial>. For any further clarifications, you may directly contact the insurer."

   iii. The above statement shall form the mandatory part of the communication to be sent to the policyholder in every case of settlement or denial of the claims.

   iv. The insurer and the TPA shall be responsible for the proper and prompt service to the policyholders at all times."
Phase 4 - Traceability Links Recovery

• Find **System Use Case – Regulations** correspondence for each system use case as a query and regulations as a database.

• One system use case may need to comply with zero or more regulations.

• Find top 5 regulations corresponding to each system use case.
Solution Approach
Phase 4 – Traceability Links Recovery

• Compute textual similarity between system use cases and regulations.
• Similarity can be purely lexical, stemmed terms of text or semantic similarity.
• SEMILAR API for similarity measurement.
• Assign each regulation a similarity score between 0 to 1 for every system use case.
• Compare performance of different techniques used.
Similarity Computation techniques

1. **Lexical**: Based on exact term matches.
2. **Meteor**: Based on alignments between SUC and regulation pair according to stemmed, synonym, and paraphrase matches.
3. **BLEU**: Based on comparison of n-grams of a sentence with the n-grams of the other.
4. **Corley**: Based on idf-weighted average of maximum similarities between each word in text 1 to any word in text 2.
5. **LDA**: Based on dot product between the corresponding vectors representing the contributions of each term to a topic.
Use Case:
In order to process the deceased claim, further requirements such as proof of death needs to be called for.

A life insurance policy shall state the primary documents which are normally required to be submitted by a claimant in support of a claim.

A claim form submitted shall be accompanied by such other documents in support of the claim as the Board may require.

In order to accept a policy which is porting-in, insurer shall not levy any additional loading or charges exclusively for the purpose of porting.

All health insurers and TPAs, as the case may be, shall establish a separate channel to address the health insurance related claims and grievances of senior citizens.

Every insurer shall inform and keep informed periodically the insured on the requirements to be fulfilled by the insured regarding lodging of a claim.

IRDA regulations, 2002
Use Case:
The **policy holder** has to **fill** in all the **particulars** of the **switch form** and **send it back** to the **insurance company** in order to **switch policy**.

The **policyholder** shall **fill** in the **portability form** along with the proposal form and **submit** the same to the **insurance company**.

On receipt of **portability form**, the **insurance company** shall address the existing insurance company seeking necessary **details** of claim history of the concerned **policyholder**.

A **policy holder** desirous of **porting his policy** to another **insurance company** shall apply to such insurance company, **to port the entire policy** along with all the members of the family, if any at **least 45 days before** the premium renewal date of his/her existing policy.

A life insurance policy shall state the primary documents which are normally required to be submitted by a claimant in support of a claim.

The insurer shall furnish to the policyholder a written acknowledgment of having registered a nomination and charge a fee not exceeding one rupee for registering such cancellation or change.
Presentation Outline

✓ Research Motivation & Aim
✓ Related work & Research Contributions
✓ Research Framework & Experimental Dataset
✓ Solution Approach
   ✓ Automated Extraction of System Use Cases
   ✓ Automated Traceability Links Recovery

• Performance Evaluation & Results
• Limitations
• Conclusion
# Performance Evaluation & Results

## System Use Case Extraction

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCTI</td>
<td>1 4 (8%)</td>
<td>39 (83%)</td>
<td>1 (2%)</td>
<td>3 (7%)</td>
<td>0.91</td>
</tr>
<tr>
<td>2</td>
<td>SUCI</td>
<td>44 (54%)</td>
<td>36 (44%)</td>
<td>2 (2%)</td>
<td>0 (0%)</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>UCPI</td>
<td>15 (83%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>3 (17%)</td>
<td>0.83</td>
</tr>
<tr>
<td>4</td>
<td>SETD</td>
<td>26 (32%)</td>
<td>7 (9%)</td>
<td>0 (0%)</td>
<td>47 (59%)</td>
<td>0.41</td>
</tr>
<tr>
<td>5</td>
<td>IDND</td>
<td>68 (85%)</td>
<td>2 (3%)</td>
<td>5 (6%)</td>
<td>5 (6%)</td>
<td>0.88</td>
</tr>
<tr>
<td>6</td>
<td>DAAD</td>
<td>30 (37%)</td>
<td>3 (4%)</td>
<td>4 (5%)</td>
<td>43 (54%)</td>
<td>0.41</td>
</tr>
<tr>
<td>7</td>
<td>UIUI</td>
<td>6 (7%)</td>
<td>7 (9%)</td>
<td>0 (0%)</td>
<td>67 (84%)</td>
<td>0.16</td>
</tr>
<tr>
<td>8</td>
<td>CUCI</td>
<td>55 (69%)</td>
<td>5 (6%)</td>
<td>2 (3%)</td>
<td>18 (22%)</td>
<td>0.75</td>
</tr>
<tr>
<td>9</td>
<td>ETUI</td>
<td>67 (84%)</td>
<td>2 (2%)</td>
<td>5 (6%)</td>
<td>6 (8%)</td>
<td>0.86</td>
</tr>
<tr>
<td>10</td>
<td>NAUI</td>
<td>71 (89%)</td>
<td>1 (1%)</td>
<td>6 (7%)</td>
<td>2 (3%)</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Feature Wise Accuracy Result
Total use cases extracted using combination of features on complete dataset were 1264.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

Accuracy of system use case extraction combining all the features together \(~83\%\)
Performance Evaluation & Results

Traceability Links Recovery

• The top 5 regulations matches are found for system use cases for three different verticals of insurance domain as shown:

<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Dataset Vertical</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPRQ</td>
<td>Claim Processing, Processing Time, Policy Premium</td>
</tr>
<tr>
<td>PRPQ</td>
<td>Policy Registration, Policy Portability, Premium payment time restriction</td>
</tr>
<tr>
<td>GHIQ</td>
<td>General Provisions of health insurance, Policy Renewals</td>
</tr>
</tbody>
</table>

Dataset Bifurcations for Compliance Detection
Performance Evaluation & Results

Traceability Links Recovery

• **Proposed Metric – Average Precision (AP)**
  
  – reflects the performance of retrieval over all relevant documents by averaging the precision value obtained after each relevant document is retrieved.
  
  – rewards systems that retrieve relevant documents quickly (highly ranked).

  \[
  AP = \frac{\sum_{k=1}^{n} P(k) \times rel(k)}{\text{Number of relevant documents}}
  \]

  – Where

  • k – the index of statement under observation
  • N - number of statements considered. Here, 5.
  • P – Precision of statement k
  • Rel – Relevance of statement k (0 or 1)
Performance Evaluation & Results

Traceability Links Recovery

For CPRQ, the AP scores are as follows:

<table>
<thead>
<tr>
<th>Average Precision</th>
<th>C1 METEOR</th>
<th>C2 Lexical</th>
<th>C3 BLEU</th>
<th>C4 CORLEY</th>
<th>C5 LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>UC1</td>
<td>0</td>
<td>0.7</td>
<td>0</td>
<td>0.42</td>
<td>0.75</td>
</tr>
<tr>
<td>UC2</td>
<td>0.25</td>
<td>0.75</td>
<td>0.35</td>
<td>0.08</td>
<td>0.69</td>
</tr>
<tr>
<td>UC3</td>
<td>0.5</td>
<td>0.25</td>
<td>0.05</td>
<td>0.4</td>
<td>0.57</td>
</tr>
<tr>
<td>UC4</td>
<td>0.5</td>
<td>0.18</td>
<td>0.18</td>
<td>0.23</td>
<td>0.38</td>
</tr>
<tr>
<td>UC5</td>
<td>0.5</td>
<td>0.88</td>
<td>0.35</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>UC6</td>
<td>0.28</td>
<td>0.67</td>
<td>0.06</td>
<td>0.56</td>
<td>0.33</td>
</tr>
<tr>
<td>UC7</td>
<td>0.33</td>
<td>0.56</td>
<td>0.47</td>
<td>0.11</td>
<td>0.91</td>
</tr>
<tr>
<td>UC8</td>
<td>0.67</td>
<td>0.7</td>
<td>0.47</td>
<td>0.56</td>
<td>0.8</td>
</tr>
<tr>
<td>UC9</td>
<td>0.67</td>
<td>0.33</td>
<td>0.07</td>
<td>0.53</td>
<td>0.75</td>
</tr>
<tr>
<td>UC10</td>
<td>0.28</td>
<td>0.5</td>
<td>0.07</td>
<td>0.56</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Average Precision Scores for Traceability Links Recovery for CPRQ
Performance Evaluation & Results

Traceability Links Recovery

For CPRQ, the AP scores graphically:

• Equally distributed colors in LDA indicates its better performance than other comparers show biased behavior.
Performance Evaluation & Results

Traceability Links Recovery

For PRPQ, the AP scores graphically:

• Meteor performance increases but LDA is still the best.
Performance Evaluation & Results

Traceability Links Recovery

• To analyze the performance of comparers over complete dataset, we propose another standard metric –
  – Mean Average Precision (MAP): Mean of the average precision scores for each query.

\[
MAP = \frac{\sum_{q=1}^{n} AP(q)}{n}
\]

• Where,
  • q - Query (System use case here).
  • AP - Average precision score for query q
  • n - Total number of system use cases.

<table>
<thead>
<tr>
<th></th>
<th>Lexical</th>
<th>Meteor</th>
<th>BLEU</th>
<th>Corley</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.48</td>
<td>0.47</td>
<td>0.24</td>
<td>0.43</td>
<td>0.72</td>
</tr>
</tbody>
</table>

[47]
Presentation Outline

✓ Research Motivation & Aim
✓ Related work & Research Contributions
✓ Research Framework & Experimental Dataset
✓ Solution Approach
  ✓ Automated Extraction of System Use Cases
  ✓ Automated Traceability Links Recovery
✓ Performance Evaluation & Results
  • Limitations
  • Conclusion
Limitations

• Dependency on basic document writing structure for certain features.
• Tables with inconsistent number of rows & columns are not detected.
• Syntactic patterns are not rigid enough and hence may result in false positives.
Presentation Outline

✓ Research Motivation & Aim
✓ Related work & Research Contributions
✓ Research Framework & Experimental Dataset
✓ Solution Approach
  ✓ Automated Extraction of System Use Cases
  ✓ Automated Traceability Links Recovery
✓ Performance Evaluation & Results
✓ Limitations
  • Conclusion
Conclusion

• System use case can be extracted automatically with good enough accuracy if basic document writing standards are followed.

• Automated traceability links recovery approach proposed by us eases the process for compliance detection by detecting top 5 regulations that are needed to be complied with a system use case.

• If a system use case does not need to comply with a regulation then it can be analyzed by reading only 5 regulations.

• The approach can further be improved by enriching the lexicons and syntactic patterns in system use case extraction.
References


References


Thank You
QUESTIONS?