A Data set of Extracted Rationale from Linux Kernel Commit Messages

Mouna Dhaouadi
DIRO, Université de Montréal, Montréal, Canada

Context

- **Problem:** Extracting rationale from natural text is useful but hard [1,2].
- **My PhD:** End-to-end rationale reconstruction and analysis [3,4,5].
- **Focus here:** Rationale in commit messages in the Linux Out-of-Memory Killer module
- **Contribution:** An annotated high quality rationale dataset

### Dataset Creation*

Manual labelling of 418 commit messages:
- Pre-processing:
  - Remove merge commits
  - Filter code sentences
- Six piloting rounds:
  - Codebook
  - A shared protocol
- Labelling by batches:
  - Sentence-based
  - Multi-label sentences
  - Three annotators
  - Discuss when conflicts
  - Fleiss Kappa ≈ 0.66

*The dataset is publicly available at: https://zenodo.org/records/10063089

### Label Meaning

<table>
<thead>
<tr>
<th>Label</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td>An action or a change that has been made, including a description of the patch behaviour</td>
</tr>
<tr>
<td>Rationale</td>
<td>Reason for a decision or value judgement</td>
</tr>
<tr>
<td>Supporting Facts</td>
<td>A narration of facts used to support a decision</td>
</tr>
<tr>
<td>Inapplicable</td>
<td>Pre-processing error or bad sentences (i.e., does not contain English sentences)</td>
</tr>
</tbody>
</table>

### Codebook

- **Decision:** The 'start' and 'end' arguments to tlb_gather_mmu() are no longer needed now that there is a separate function for 'fullmm flushing'
- **Rationale:** Remove the unused arguments and update all callers.

### Supporting Facts

- **Rationale:** This patch introduces printk ratelimiting specifically for the oom killer
- **Supporting Facts:** The oom killer should not be subjected to this state just because another subsystem or driver may be flooding the kernel log

### Dataset Analysis

- **Almost all** the commits contain rationale information
- **In average,** 60% of the commit message contains rationale information
- The quantity of rationale reported depends neither on the commit message size nor the developers’ experience.
- As a commit becomes longer, it tends to have 40% to 60% of its sentences containing rationale
- **Experienced** developers have a rationale density around 60%

### References


### Contact

Mouna Dhaouadi
Université de Montréal
Email: mouna.dhaouadi@umontreal.ca
Website: https://www.labs.lro.umontreal.ca/~dhaouadm/

This work is partially supported by the ACM-W scholarship program and Google.

---

**Problem:** Extracting rationale from natural text is useful but hard [1,2].

**My PhD:** End-to-end rationale reconstruction and analysis [3,4,5].

**Focus here:** Rationale in commit messages in the Linux Out-of-Memory Killer module

**Contribution:** An annotated high quality rationale dataset

---

**Context**

- **Problem:** Extracting rationale from natural text is useful but hard [1,2].
- **My PhD:** End-to-end rationale reconstruction and analysis [3,4,5].
- **Focus here:** Rationale in commit messages in the Linux Out-of-Memory Killer module
- **Contribution:** An annotated high quality rationale dataset

### Dataset Creation*

Manual labelling of 418 commit messages:
- Pre-processing:
  - Remove merge commits
  - Filter code sentences
- Six piloting rounds:
  - Codebook
  - A shared protocol
- Labelling by batches:
  - Sentence-based
  - Multi-label sentences
  - Three annotators
  - Discuss when conflicts
  - Fleiss Kappa ≈ 0.66

*The dataset is publicly available at: https://zenodo.org/records/10063089

### Label Meaning

<table>
<thead>
<tr>
<th>Label</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td>An action or a change that has been made, including a description of the patch behaviour</td>
</tr>
<tr>
<td>Rationale</td>
<td>Reason for a decision or value judgement</td>
</tr>
<tr>
<td>Supporting Facts</td>
<td>A narration of facts used to support a decision</td>
</tr>
<tr>
<td>Inapplicable</td>
<td>Pre-processing error or bad sentences (i.e., does not contain English sentences)</td>
</tr>
</tbody>
</table>

### Codebook

- **Decision:** The 'start' and 'end' arguments to tlb_gather_mmu() are no longer needed now that there is a separate function for 'fullmm flushing'
- **Rationale:** Remove the unused arguments and update all callers.

### Supporting Facts

- **Rationale:** This patch introduces printk ratelimiting specifically for the oom killer
- **Supporting Facts:** The oom killer should not be subjected to this state just because another subsystem or driver may be flooding the kernel log

### Dataset Analysis

- **Almost all** the commits contain rationale information
- **In average,** 60% of the commit message contains rationale information
- The quantity of rationale reported depends neither on the commit message size nor the developers’ experience.
- As a commit becomes longer, it tends to have 40% to 60% of its sentences containing rationale
- **Experienced** developers have a rationale density around 60%

---

**References**


---

**Contact**

Mouna Dhaouadi
Université de Montréal
Email: mouna.dhaouadi@umontreal.ca
Website: https://www.labs.lro.umontreal.ca/~dhaouadm/

This work is partially supported by the ACM-W scholarship program and Google.