A Data Set of Extracted Rationale from Linux Kernel Commit Messages

Mouna Dhaouadi
Université de Montréal
Montreal, Canada
mouna.dhaouadi@umontreal.ca

ABSTRACT
Developer’s commit messages contain information about decisions taken and their rationale. Extracting this information is challenging since we lack a detailed understanding of how developers express these concepts. Our work-in-progress targets this challenge by producing a labelled data set of commit messages for a Linux Kernel component. We report preliminary analyses which suggest that larger commit messages and more experienced developers commits tend towards having 40% of sentences containing rationale. This may indicate a guideline for developers to target.

CCS CONCEPTS
- Software and its engineering → Software design engineering.

KEYWORDS
Software rationale, Linux kernel, Data set, Commit messages

ACM Reference Format:

1 RESEARCH PROBLEM AND MOTIVATION
Rationale for decisions and changes in software projects is often recorded in the commit messages [1]. Capturing this implicit knowledge is useful to create a record that could be used when revisiting decisions or to produce future recommendations [2].

The Linux kernel is a large-scale open-source project involving many collaborators. Thus, having a shared understanding is necessary to make coherent decisions. Linux kernel commit messages usually contain a description of the rationale behind the introduced changes. However, this rationale information is embedded inside the commits and we lack a systematic process of capturing and organizing decisions and their rationale.

Our prior work has proposed a vision for an end-to-end reconstruction pipeline to explicitly structure this rationale and relate it to the decisions [3]. In this ongoing work, we create a high-quality data set from a subset of Linux kernel commits. This data set will help us develop our pipeline to improve the developer’s knowledge of the code base, and better understand how decisions and rationale are recorded.

The two main contributions of this paper are: 1) a labelled data set of 1144 sentences extracted from 160 commits from the Out-Of-Memory Killer (OOM-Killer) kernel component, and 2) initial analyses and insights concerning the abundance (RQ1), amount (RQ2) and developer experience (RQ3) characteristics of rationale.

2 BACKGROUND AND RELATED WORK
Developers rationale refers to the reasoning behind the decisions that developers make [2]. Prior work has tried to extract rationale from textual artifacts for different projects; e.g., from Python email archives [6], from Apache project commit messages [4] and from Chrome Bug reports [5]. However, to the best of our knowledge, there is no previous work that proposed a data set of extracted rationale from the commit messages for the Linux kernel. In particular, Linux developers are encouraged to describe concisely their rationale/motivation in the commit descriptions1. This makes Linux commit messages a comprehensive and very semantically-rich repository of decision/rationale information.

3 APPROACH AND CONTRIBUTIONS
To create our data set, we obtained the commit history (418 commits) of the OOM-killer file2. For each commit, we reduced noise (e.g., removed code and meta data), and split it into sentences to analyze.

Three annotators (including myself) iterated over six piloting rounds to reach a consensus regarding the set of labels to use (Table 1). We agreed to consider terse value judgment language (e.g., “fix” or “cleanup”) to imply the presence of rationale and decisions. For instance, the sentence “fix it” is considered both

<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 judgment</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>

1https://www.kernel.org/doc/html/v4.10/process/submitting-patches.html#describe-your-changes
2https://github.com/torvalds/linux/commits/master/mm/oom_kill.c accessed on 12/01/2023
Decision and Rationale. We provide an example of labelling a commit message\(^3\) in Table 2.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>signal: Use SEND_SIG_PRIV not SEND_SIG_FORCED with SIGKILL and SIGSTOP</td>
<td>Decision</td>
</tr>
<tr>
<td>Now that siginfo is never allocated for SIGKILL and SIGSTOP there is no difference between SEND_SIG_PRIV and SEND_SIG_FORCED for SIGKILL and SIGSTOP.</td>
<td>Supporting Facts</td>
</tr>
<tr>
<td>This makes SEND_SIG_FORCED unnecessary and redundant in the presence of SIGKILL and SIGSTOP.</td>
<td>Rationale</td>
</tr>
<tr>
<td>Therefore change users of SEND_SIG_FORCED that are sending SIGKILL or SIGSTOP to use SEND_SIG_PRIV instead.</td>
<td>Decision</td>
</tr>
<tr>
<td>This removes the last users of SEND_SIG_FORCED.</td>
<td></td>
</tr>
</tbody>
</table>

During the labelling process, Fleiss Kappa averaged 0.68 for seven rounds (so far). This indicates strong agreement considering the subjective nature of rationale \(^2\).

**RQ1. How many commits contain rationale?** 97.5% of the commits contain at least one sentence with rationale information. This suggests that rationale is almost always described.

**RQ2. How much of the commit contains rationale?** The rationale density is the percentage of commit sentences that contain rationale. Figure 1a shows these values versus the size of commits. The figure shows that as a commit becomes longer, it tends to have 40% of its sentences contain rationale information. Overall, 43.87% of all the commit messages contains rationale information.

**RQ3. Does the quantity of rationale reported depend on the developer experience?** We visualize the average rationale density per author in Figure 1b. Most of developers commits have a density between 30% and 60%, but more experienced developers commits have a density near 40%.

## ACKNOWLEDGMENTS

Partially funded by the Fonds de Recherche du Québec (B2X). I also thank the other project members: Bentley James Oakes and Michalis Famelis.

## REFERENCES


Received 2023-06-05; accepted 2023-08-11