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*

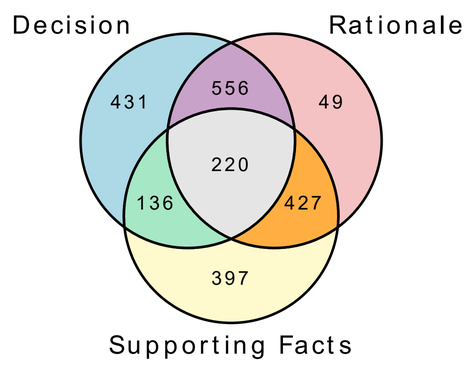
Manual labelling of 418 commit messages:

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

Codebook



Distribution of the labelled sentences

Sentence	Labelling
mm, oom: introduce independent oom killer ratelimit state	Decision
printk_ratelimit() uses the global ratelimit state for all printks	Supporting Facts
The oom killer should not be subjected to this state just because another subsystem or driver may be flooding the kernel log	Rationale
This patch introduces printk ratelimiting specifically for the oom killer	Decision

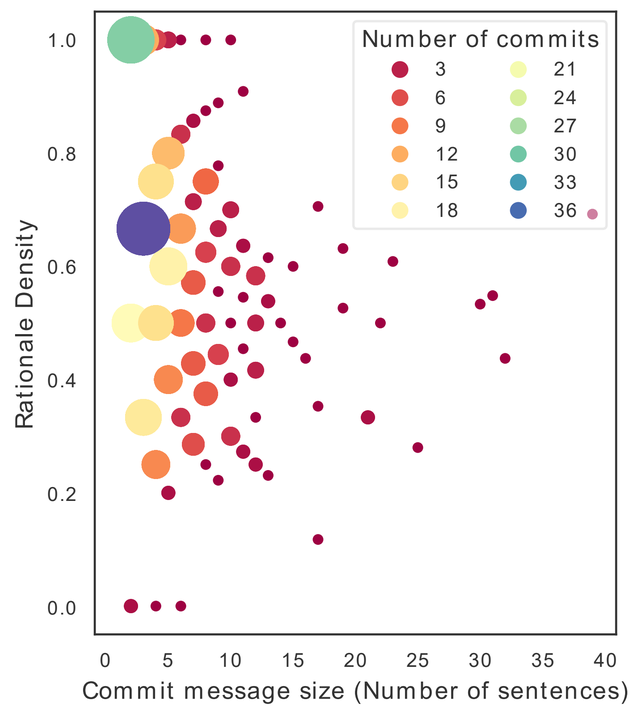
An example commit with labelled sentences from our dataset

Sentence	Labelling
tlb: mmu_gather: Remove start/end arguments from tlb_gather_mmu()	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, Supporting Facts
Remove the unused arguments and update all callers.	Decision, Rationale

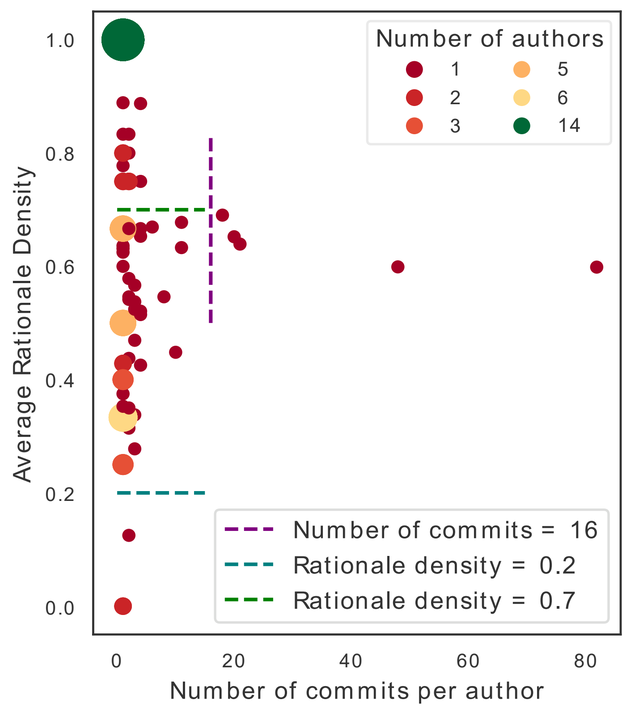
An example commit with multi-labelled sentences from our dataset

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%**



Commit size versus rationale density



Commits per author versus average rationale density

Contact

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

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

References

1. J. E. Burge, J. M. Carroll, R. McCall, and I. Mistrik, ‘’Rationale-based software engineering’’, Springer, 2008.
2. Y. Liu, Y. Liang, C. K. Kwong, and W. B. Lee, “A new design rationale representation model for rationale mining”, Journal of Computing and Information Science in Engineering, vol. 10, no. 3, 2010.
3. M. Dhaouadi, B. J. Oakes, and M. Famelis, “End-to-end rationale reconstruction”, in 37th IEEE/ACM International Conference on Automated Software Engineering, 2022, pp. 1–5
4. M. Dhaouadi, ‘’Extraction and Management of Rationale’’, in 37th IEEE/ACM International Conference on Automated Software Engineering, 2022.
5. M. Dhaouadi, B. J. Oakes, and M. Famelis, “Towards Understanding and Analyzing Rationale in Commit Messages using a Knowledge Graph Approach”, in Model Driven Engineering Languages and Systems (MODELS) Companion, 2023.