Investigating mislabeling in making a dataset for automated NFR classification

Yuta Kimura (B4)
kimura.yuta.1@g.wakayama-u.jp

Masao Ohira
masao@wakayama-u.ac.jp

wakayama univ.

Background

- Requirement specifications roughly consist of :
  - Functional requirements (FRs) :
    functions required for a system
  - Non-functional requirements (NFRs) :
    system's properties (e.g., security and usability)
- **NFRs is likely to be much overlooked than FRs, Because NFRs are**
  - described vaguely in natural language
  - expressed differently from one developer to the next and
  - sometimes included in FRs

➡ **It is difficult for developers to identify NFRs comprehensively.**

In order to help developers identify NFRs, automated NFR classification methods have been proposed.

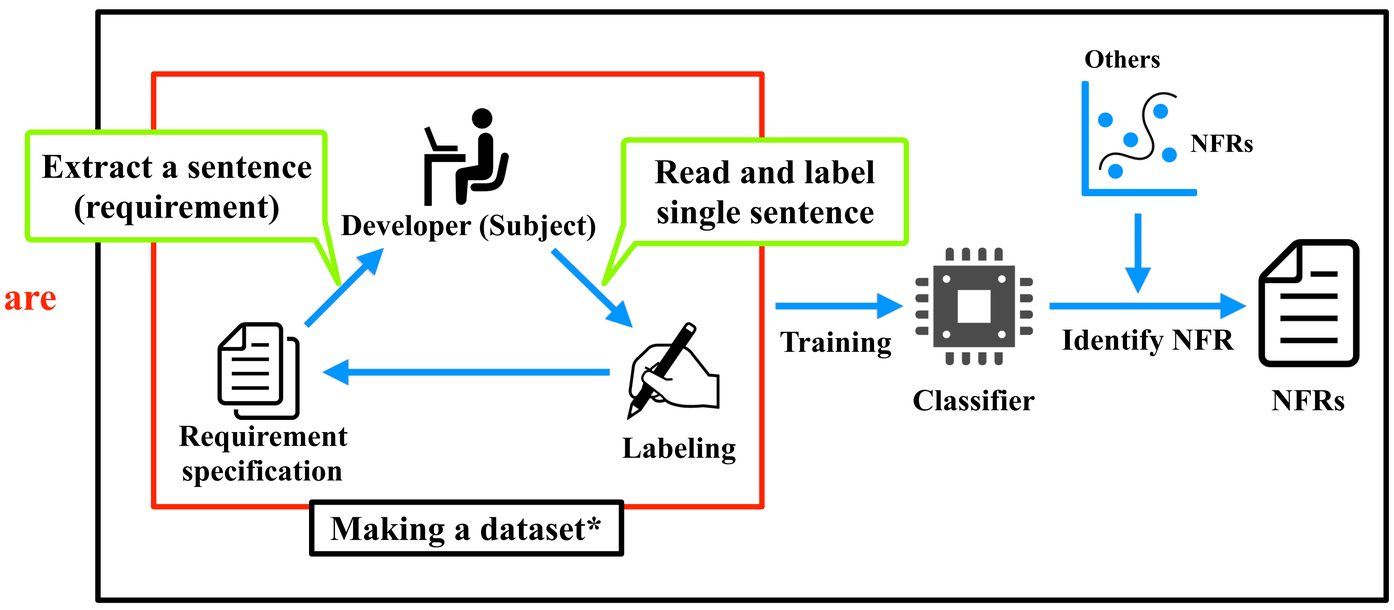

Figure 1: Existing automated NFR classification

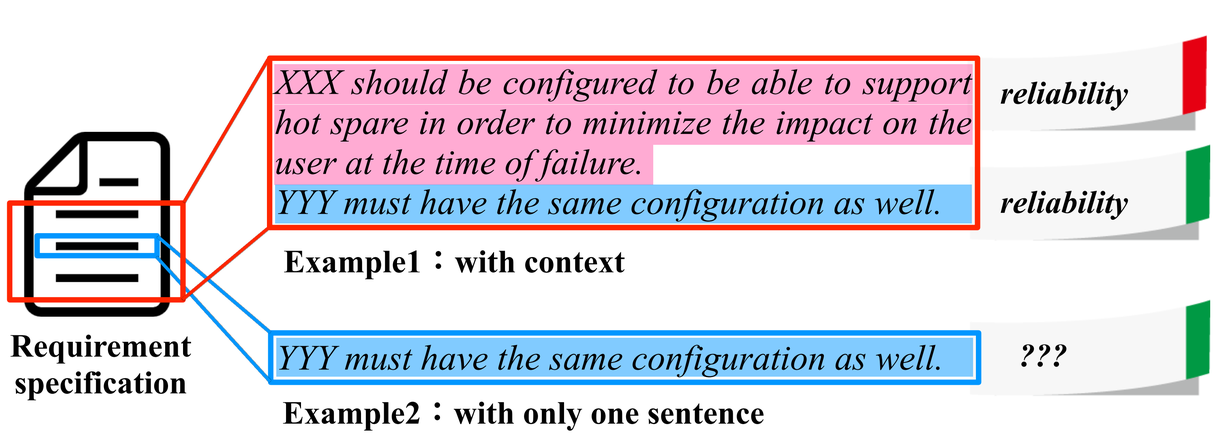

Figure 2: Examples of requirements

For Exp.1, the red highlighted requirement is labeled as reliability and the blue highlighted requirement indicates reliability. For Exp.2, however, the blue highlighted requirement is unknown.

Motivation and Aim

- **(Motivation)**
  - A dataset for existing automated NFR classification methods is created by **reading and labeling each single sentence (Fig.1 *)**.
    ➡ We suspect that requirements are not independent from the context.
  - **Is there a difference between results when labeling with single sentence and when labeling with context ?**
- **(Aim)**

We would like to reveal whether the labeling processes has an impact on the labeling result.

Experiment

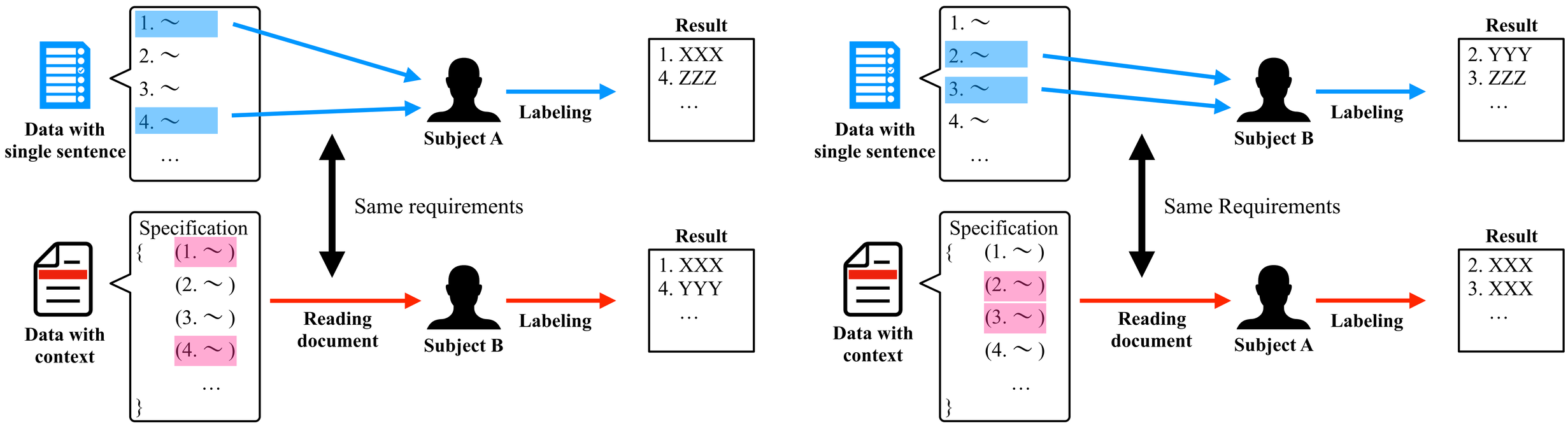

Figure 3: Example of user study

- We conduct a user study on requirement labelings.
  - Data with single sentence :
    A set of randomly extracted sentences like Exp.1 in Fig.2
  - Data with context :
    A set of highlighted sentences with context like Exp.2 in Fig.2
- Conducting two patterns to balance the learning effect (Fig.3).
  - Data with single sentence → Data with context … A
  - Data with context → Data with single sentence … B
- Verifying whether there is a statistically significant difference between A and B.

Future Work

- Conducting a user study and revealing whether there is a difference or not.
  - If there is a difference:
    ➡ We create and evaluate a classifier using a dataset with context.
  - If there is not a difference:
    ➡ It is regarded that a dataset creation process for existing methods have no problem.