Evaluation of Example Tools For Hairy Tasks

Presenter:
Changsheng chen

CS 846 project Presentation
Department of Computer Science
Outline

▪ Motivation
▪ Introduction
▪ Related works
▪ Case study 1
▪ Case study 2
▪ Conclusion
Motivation

- In some scenarios, for some tasks, any tool with less than 100% recall is not helpful and the user may be better off doing the task entirely manually.
- The trade off between precision and recall may make it difficult to interpret the true result.
- Improper use of precision and recall may affect evaluation.
- Different tasks need different weight for F-measure
Recall (R) is the percentage of the correct answers that the tool returns:

\[ R = \frac{|\text{ret} \cap \text{cor}|}{|\text{cor}|} = \frac{|TP|}{|TP| + |FN|} \]

Which is the percentage of the right stuff that is found.

Precision (P) is the percentage of the tool-returned answers that are correct:

\[ P = \frac{|\text{ret} \cap \text{cor}|}{|\text{ret}|} = \frac{|TP|}{|FP| + |TP|} \]

Precision is the percentage of the found stuff that is right.
Introduction – F-Measure

- F-measure: harmonic mean of Precision and Recall

\[ F_1 = \left( \frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \]

- Weighted F-Measure: For situations in which R and P are not equally important.

\[ F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}. \]

\( \beta \) is the ratio by which it is desired to weight Recall more than Precision.
Case Study 1:

- Using Tools to Assist Identification of Non-requirements in Requirements Specifications – A Controlled Experiment (Jonas Paul Winkler and Andreas Vogelsang)
  - Categorizing textual fragments into requirements and non-requirements.
  - In practice, this categorization is performed manually.
  - Developed a tool to assist users in this task by providing warnings based on classification.
  - Performed a controlled experiment with two groups of students.
  - The results show that the application of an automated classification approach may provide benefits, given that the accuracy is high enough.
Case Study 1:

- Using Tools to Assist Identification of Non-requirements in Requirements Specifications – A Controlled Experiment (Jonas Paul Winkler and Andreas Vogelsang)

- Investigation of the effectiveness of automated tools for RE tasks
  - Their experiment supports that claim that the accuracy of the tool may have an effect on the observed performance.
  - A human working with the tool on the task should at least achieve better recall than a human working on the task entirely manually.
  - The experimental setup follows this idea by comparing tool-assisted and manual reviews.
Case Study 2:

- Evaluation of Techniques to Detect Wrong Interaction Based Trace Links (Paul Hubner and Barbara Paech)
  - Trace links are created and used continuously during the development
  - Support developers with an automatic trace link creation approach with high precision.
  - In their previous study we showed an interaction based trace link creation approach which is better than traditional IR based approaches. Performed a controlled experiment with two groups of students.
  - Performed the study within a student project.
  - Evaluated different techniques to identify relevant trace link candidates such as focus on edit interactions or thresholds for frequency and duration of trace link candidates.
Case Study 2:

- Evaluation of Techniques to Detect Wrong Interaction Based Trace Links (Paul Hubner and Barbara Paech)
  - Trace links are created and used continuously during the development
  - Support developers with an automatic trace link creation approach with high precision.
  - In their previous study we showed an interaction based trace link creation approach which is better than traditional IR based approaches. Performed a controlled experiment with two groups of students.
  - Performed the study within a student project.
  -Evaluated different techniques to identify relevant trace link candidates such as focus on edit interactions or thresholds for frequency and duration of trace link candidates.
Conclusion

- Most RE and SE tasks involving NL documents are hairy tasks and need tools support.

- We may evaluate these tools with the different F-measure because the importance of recall and precision may be different for different tasks.

- We must to research and understand which measures are appropriate to evaluate any tool for the task.