Comparing the Accuracy of Natural Language Processing (NLP) Tools' Annotations of User Stories

Sathurshan Arulmohan[1], Sebastien Mosser PhD[1], Marie-Jean Meurs PhD[2]

[1] Department of Computing and Software, McMaster University, Hamilton, Canada. [2] Department of Computer Science, Université du Québec à Montréal, Montreal, Canada.

Agile Software Development

- Agile Software Development consists of a feedback loop, *Fig 1,* [1]
  - Product owner receives feedback from end-users and converts it into **user stories** to store in the **product backlog**
  - Developers (Dev) team uses feedback from the operations (Ops) team and from user stories in the product backlog to improve the product
- The problem is that the product backlog does not yet provide immediate feedback to Dev. team [1]
  - Slows down the feedback loop
  - Overlapping or similar feedback may not be considered at once
- An approach uses **NLP tools** to **automatically extract valuable information** from user stories to shorten the feedback loop [1]

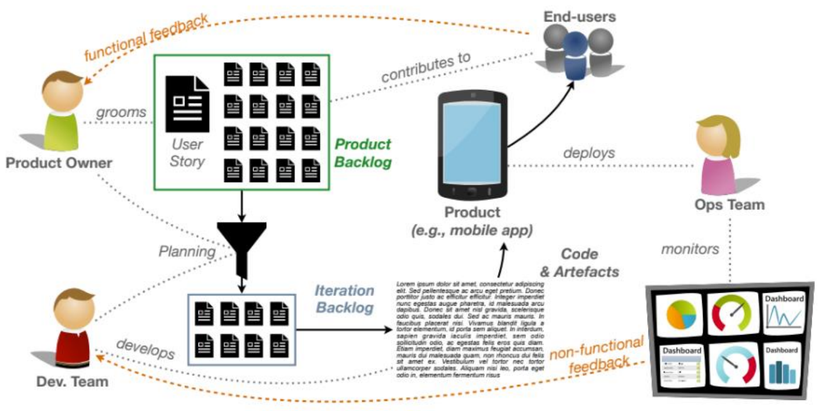

Figure 1: Agile software development feedback loop for a product. [1]

Objectives

- Evaluate the **accuracy** of current NLP tools' annotations of user stories
- Compare the accuracy of all NLP tool's annotations using a **benchmark**
  - Manually annotate each user story
- Implement a new NLP tool (**CRF** [2]) that will be more accurate at annotating than the current NLP tools

User Stories

- The only publicly available and reusable dataset is published by Dalpiaz [1],[3]
  - Consist of **22 backlogs** with **1670 unique valid user stories**
- Each user story contains a similar format [1]:

  "<PID>, As a <Persona>, I want to <perform action on entities>, so that <benefit>."

Note: some user stories are poorly written and may not have any entities or a benefit.

Annotations

Benchmark (Baseline Annotations)

- Each story was manually annotated on Doccano [4]
  - Ensures that a benchmark exists for comparing the accuracy of NLP tools' annotations
- Annotations include:
  - Labels:

    PID: Project ID

    Persona: The main person of the story

    Action: An action done by the Persona or an Entity

    Entity: Word(s) that represents an element

    Benefit: The outcome of the primary action

    Note: **Qualifiers** such as adjectives are included in Action and Entity annotations
  - Relations (between two labels):

    triggers: relation of a Persona triggering an Action

    targets: relation of Action targeting an Entity

    contains: relation of an Entity containing another Entity

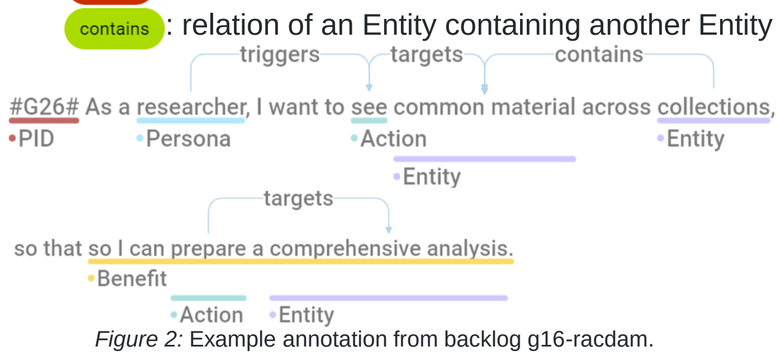

Figure 2: Example annotation from backlog g16-racdam.

- Actions and Entities are further categorized
  - **Primary**: Main Action/Entity of the story
  - **Secondary**: All other Action/Entity in the story

NLP Annotations

- **Simple NLP**: Developed a very simple annotation tool that depends on a dictionary of words
  - Used to determine if other NLP tools are redundant
- **ECMFA-VN**: Annotations of stories were already given
- **Visual Narrator**: Blackbox tool that outputs only primary annotations [5]
- **CRF**: An updated version of sklearn-crfsuite that learns from a pre-annotated training set of stories
  - Relies on pre-set **features** and **parameters** that affect its learning

Comparing Modes

- Three modes of comparison
  - **Strict**: Must EXACTLY match baseline annotations
  - **Inclusion**: Baseline results are part of NLP's results
  - **Relaxed**: Qualifiers within annotations are ignored

Example Comparison Results

Baseline Annotation	NLP Tool Annotation	Strict Comparison	Inclusion Comparison	Relaxed Comparison
Dataset	datasets	Fail	Pass	Fail
Many datasets	datasets	Fail	Fail	Pass
User's dataset	[User, datasets]	Fail	Fail	Fail
dataset	Dataset	Pass	Pass	Pass

Results

CRF Groupings Benchmark Comparison

Figure 3: Comparing the CRF results of different groupings of the training set

- Different groupings of the **training sets** showed no significant changes to the **F-Measure**
- Groups of small training sets sometimes performed worse than using one global training set

Average Benchmark Comparison

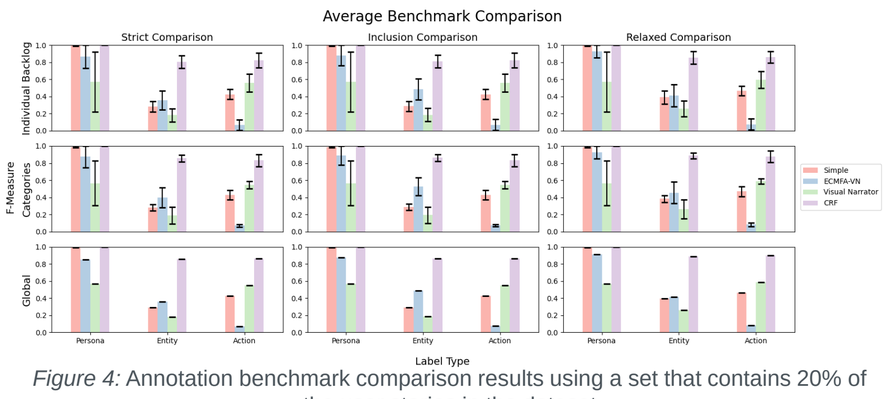

Figure 4: Annotation benchmark comparison results using a set that contains 20% of the user stories in the dataset.

Primary Results Average Benchmark Comparison

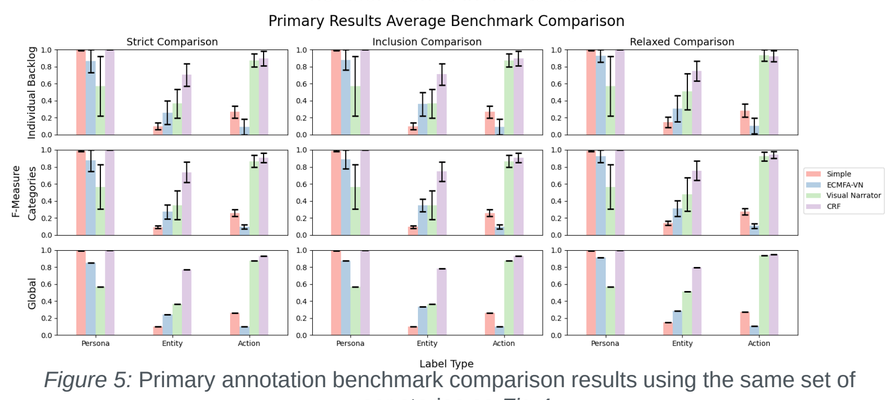

Figure 5: Primary annotation benchmark comparison results using the same set of user stories as *Fig 4.*

- ECMFA-VN does not annotate Actions well
- Visual Narrator annotates Primary Actions well but falls when annotating Personas and Entities
- CRF annotates Actions, Entities, and Primary Entities better than all the other NLP tools
- CRF can almost annotate **all** Personas and Primary Actions in a given set

Conclusions

- CRF annotates user stories well compared to existing NLP tools
  - Still has room for further improvement
- Train CRF with only POS tags when training set is large
  - Larger training sets have repeated words
  - To scale CRF, we want to avoid word dependencies
- ECMFA-VN and Visual Narrator's calculations are redundant in most cases

Future Work

- Evaluate CRF performance with different ratios of training and testing set sizes
- Evaluate CRF trained models on new datasets
- Improve CRF's relation annotations using syntactic trees and proximity matching
  - Identify Contains and secondary relations
- Evaluate the accuracy of NLP tools' relation annotations with the benchmark

References

[1] S. Mosser, C. Pulgar, V. Reinhar, "Modelling Agile Backlogs as Composable Artifacts to support Developers and Product Owners", *Journal of Object Technology*, Volume 21, no. 3, July 2022, pp. 3:1-15, doi:10.5381/jot.2022.21.3.a3.

[2] S. Mosser, "ace-sklearn-crfsuite: Scikit-learn inspired API for CRFsuite." https://github.com/ace-design/ace-sklearn-crfsuite [accessed July 29, 2022]

[3] F. Dalpiaz, "Requirements data sets (user stories)", Mendeley Data, V1, July 2018, doi: 10.17632/7zbk8zsd8y.1 [accessed May 02, 2022]

[4] H. Nakayama et al., "Doccano: Open source annotation tool for machine learning practitioners." https://github.com/doccano/doccano [accessed May 00, 2022]

[5] MJ Robeer, S.Mosser, et al. "Visual Narrator" https://github.com/MarcelRobeer/VisualNarrator [accessed May 02, 2022]

[6] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C.D. Manning. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." 2020. https:// stanfordnlp.github.io/stanza/. [accessed June 10, 2022]

Sathurshan Arulmohan
Undergraduate Software Engineering

McMaster University, Department of Computing and Software, Faculty of Engineering

Email: arulmohs@mcmaster.ca

BRIGHTER WORLD