Bachelor‘s Thesis: Conceptualization and Implementation of a Rule-based Workbench for Textual Pattern Annotation

Georg Bonczek, 2017

Chair of Software Engineering for Business Information Systems (sebis)
Faculty of Informatics
Technische Universität München
wwwmatthes.in.tum.de

Administrative Setup

- Title: **Conceptualization and Implementation of a Rule-based Workbench for Textual Pattern Annotation**

- Start: 15.08.2017
- End: 15.12.2017
- Author: Georg Bonczek (georg.bonczek@tum.de)
- Advisor: M.Sc. Bernhard Waltl (b.waltl@tum.de)

Rule-based Text Annotation

- Annotations are metadata for a span of text
- Rules consist of patterns and actions
- Patterns are RegEx like formulations for sequences of annotations

Rule-based Text Annotation

- Annotations are metadata for a span of text
- Rules consist of patterns and actions
- Patterns are RegEx like formulations for sequences of annotations

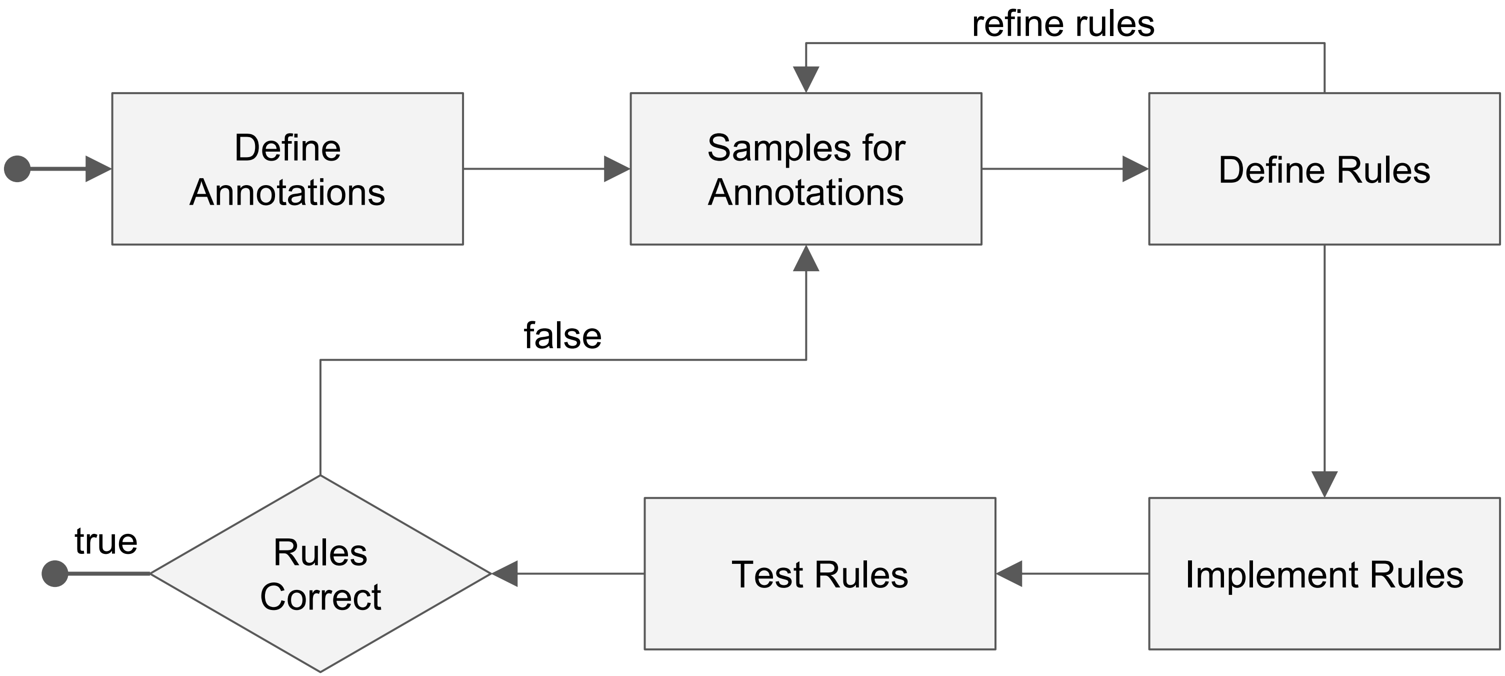
Ein Produkt hat einen Fehler, wenn...

Motivation

Rule-based text annotation is still useful in times of machine learning:

- Predictable results
- Easy and fast to implement
- Incorporation of domain knowledge
- Creation of training sets

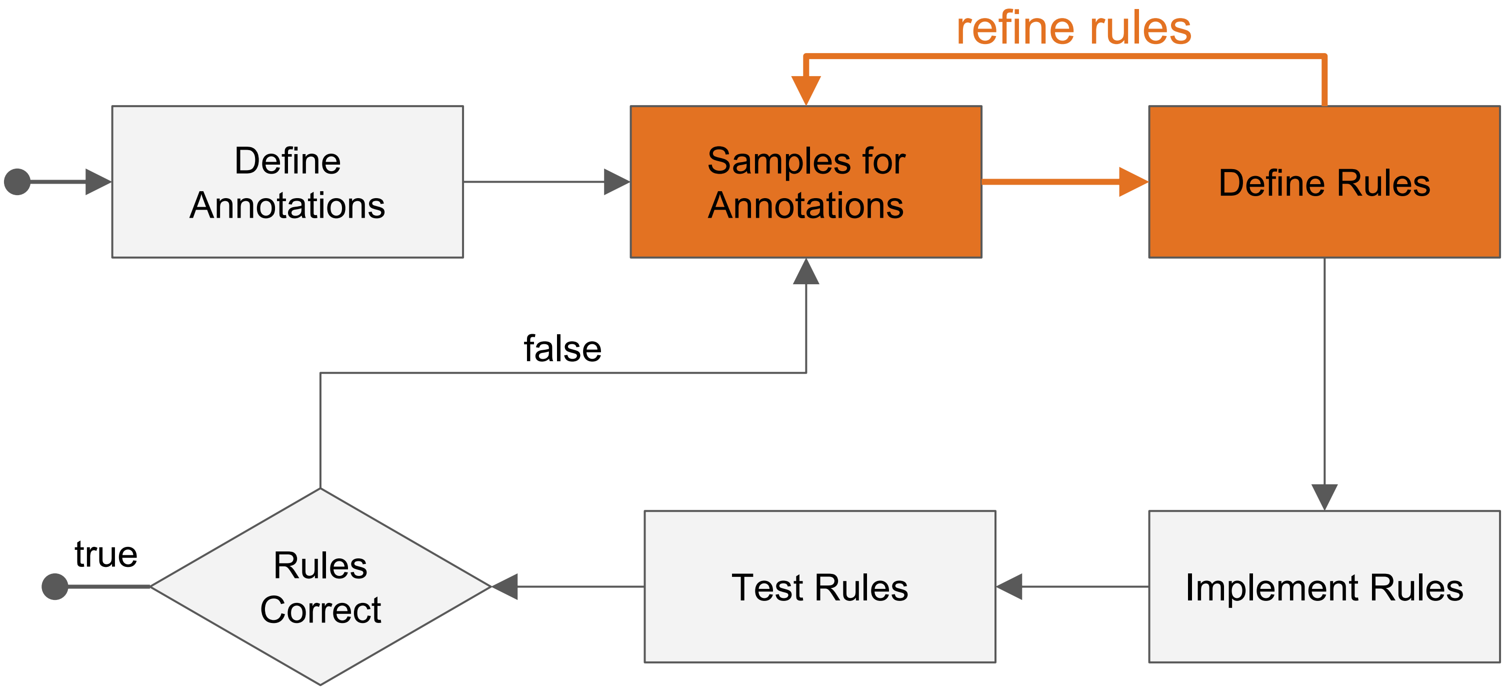
Current workflow



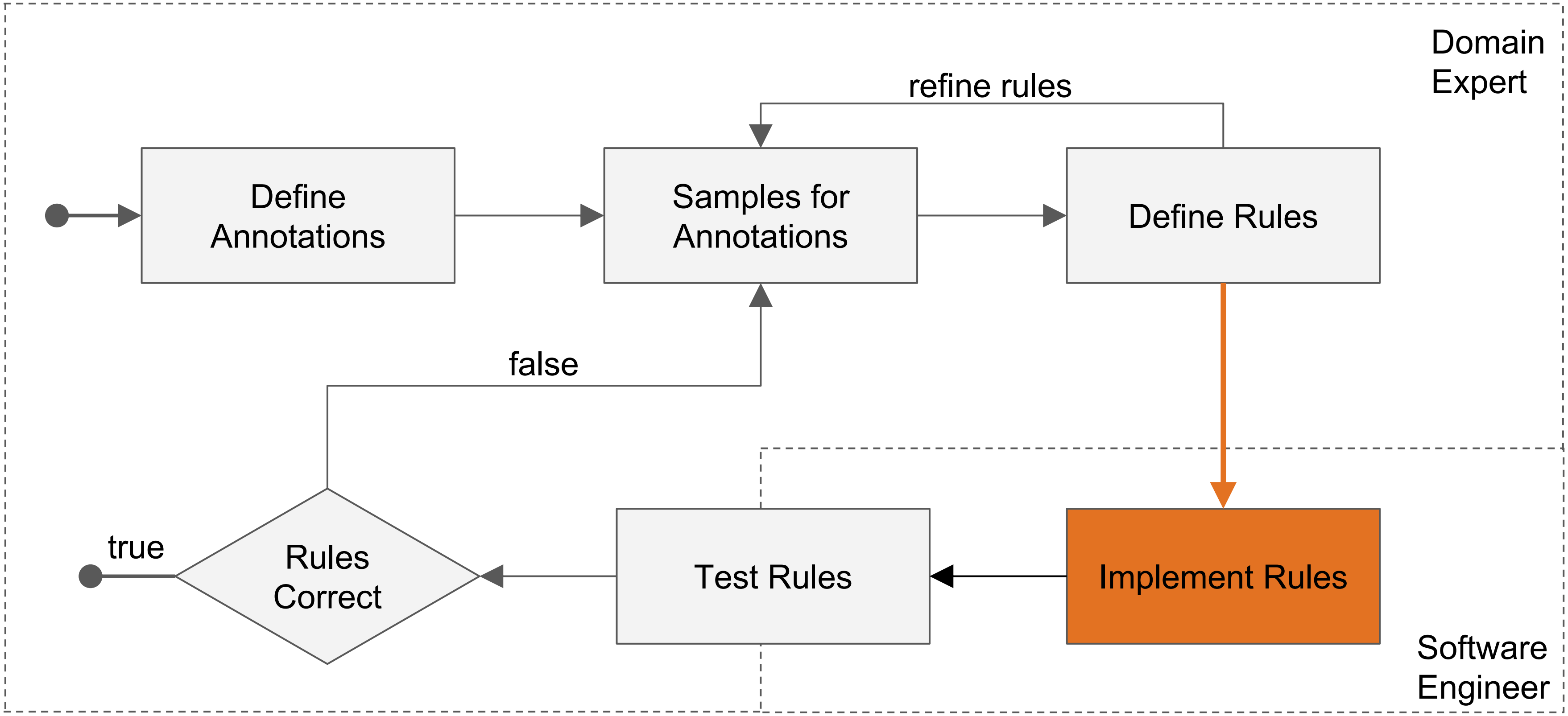
Status Quo

	GATE / JAPE IDE	UIMA / UIMA Ruta IDE
Conceptualization	X	X
Implementation	✓	✓
Testing	✓	✓
Embeddable IDE	X	X
Doesn't require technical knowledge	X	X

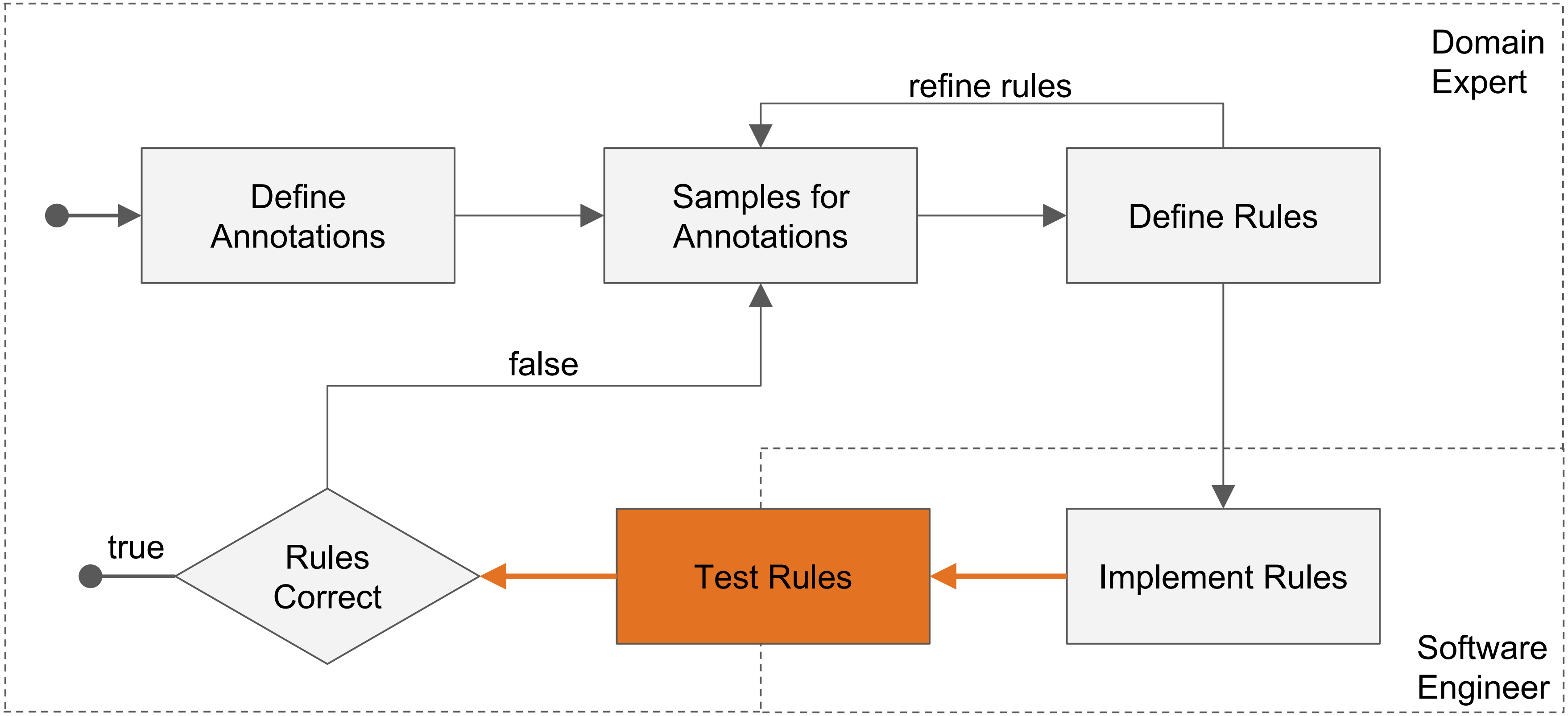
Manual Collection of Samples



Implementation Requires Communication

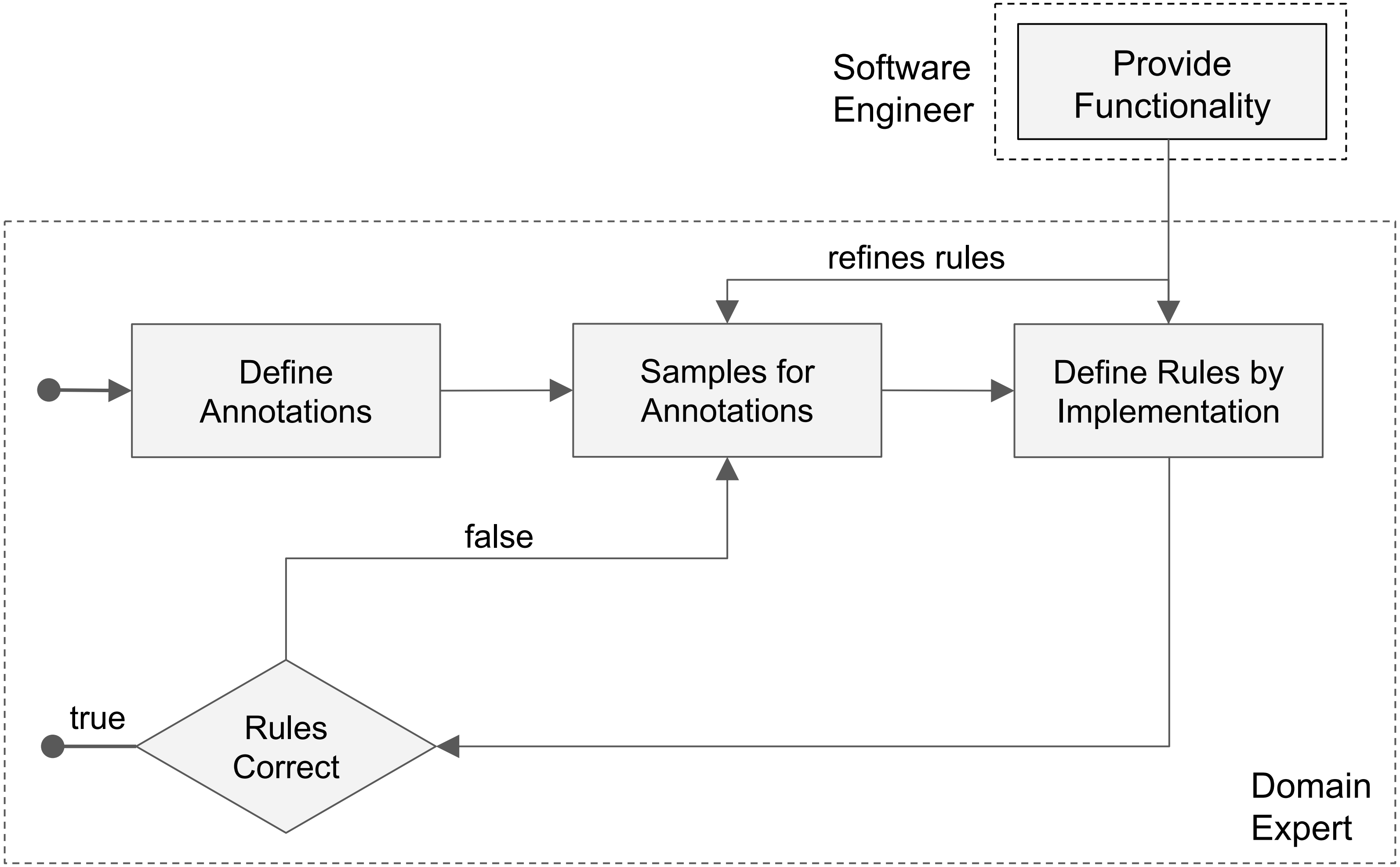


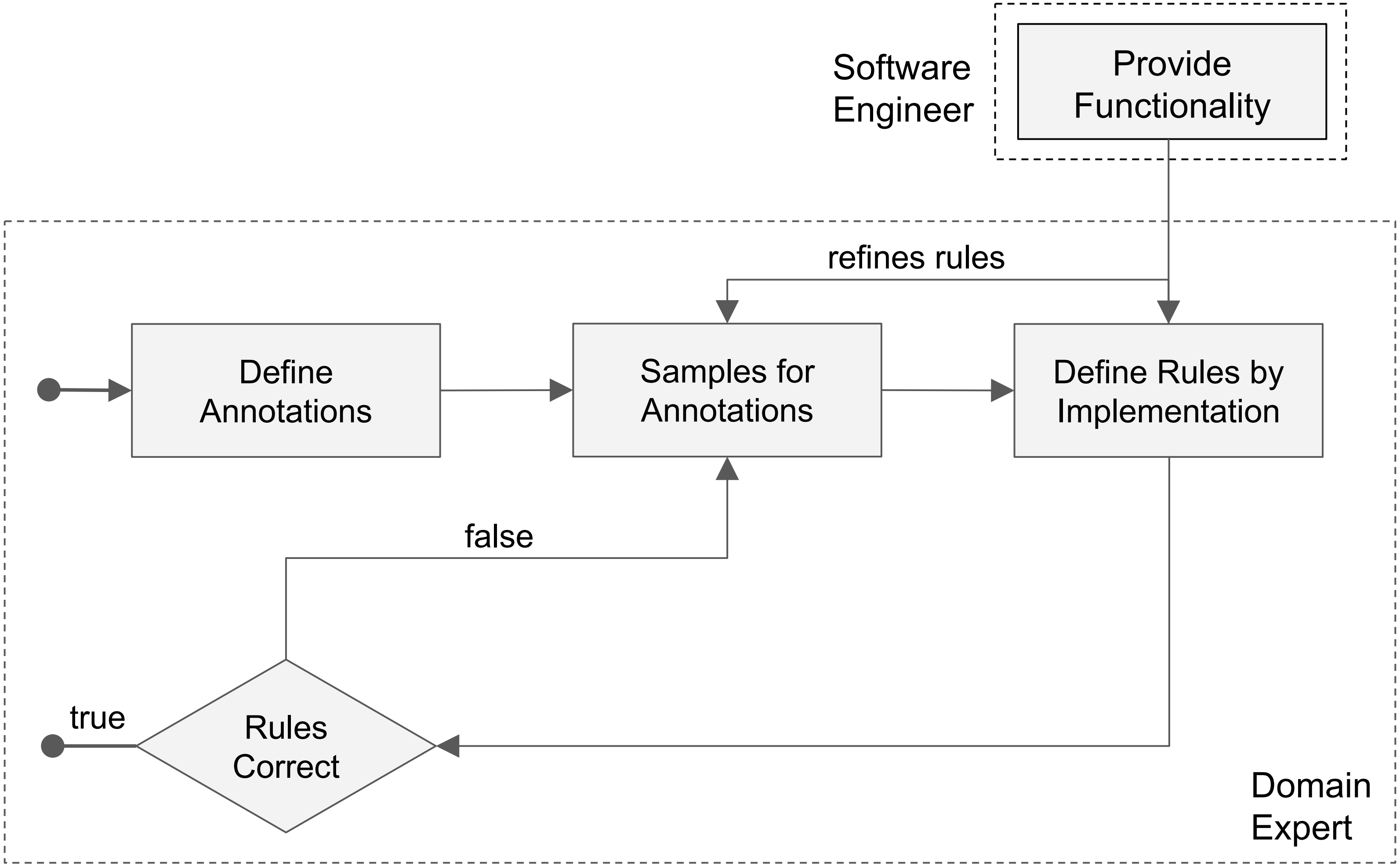
Testing Needs To Be Synchronized



Problem Statement

- Rule-based text annotation
- Current environments do not cover complete development process
- Unsuitable for non-technical domain experts
- No focus on interdisciplinary collaboration of domain experts and SE


Software Engineer
Provide Functionality
refines rules
Define Annotations
Samples for Annotations
Define Rules by Implementation
false
true
Rules Correct
Domain Expert


Software Engineer
Provide Functionality
refines rules
Define Annotations
Samples for Annotations
Define Rules by Implementation
false
true
Rules Correct
Domain Expert

Solution

- Dedicated user interfaces for the conceptualization of rules
    - Sample collection by text highlighting
    - Remove immediate need for SE
- Support rule implementation
    - Different approaches to rule editors
    - Automatic rule learning
    - …
- Automate manual tasks like testing

Research Questions

- What are the concrete **phases** in rule development?
- How can we **support** this development process?
- Which **existing technologies** can be integrated?
- How can we **separate concerns**?

Questions

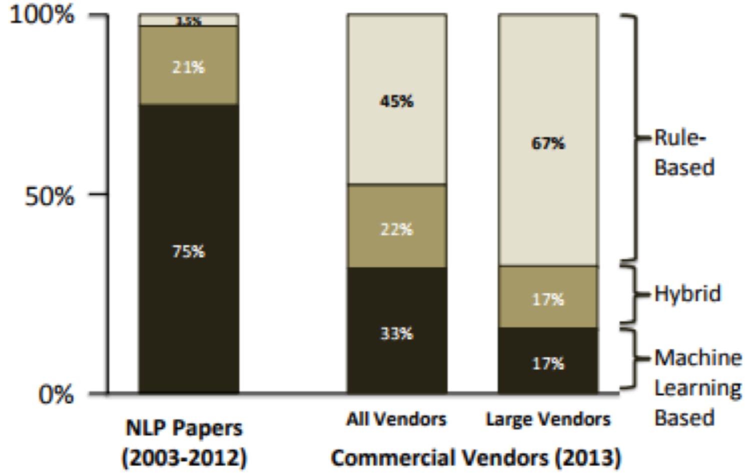
References

Figure p. 19: Chiticariu, Laura, Yunyao Li, and Frederick R. Reiss. "Rule-based information extraction is dead! long live rule-based information extraction systems!." EMNLP. No. October. 2013.

```
Phase: UrlPre
Input:  Token SpaceToken
Options: control = appelt

Rule: Urlpre

( (({Token.string == "http"} |
  {Token.string == "ftp"})
 {Token.string == ":"}
 {Token.string == "/"}
        {Token.string == "/"}
       ) |
({Token.string == "www"}
        {Token.string == "."}
       )
):urlpre
-->
:urlpre.UrlPre = {rule = "UrlPre"}
```

```
WORDLIST FirstNameList = 'FirstNames.txt';
DECLARE FirstName, FirstNameInitial, Name, NameListPart;

Document{-> MARKFAST(FirstName, FirstNameList)};

DECLARE NameLinker;
W{REGEXP("and", false) -> MARK(NameLinker)};
COMMA{ -> MARK(NameLinker)};
SPECIAL{REGEXP("&") -> MARK(NameLinker)};

CW{REGEXP(".") -> MARK(FirstNameInitial,1,2)} PERIOD;

FirstName+ FirstNameInitial* CW{-> MARK(Name, 1, 2, 3)};
FirstNameInitial+{-PARTOF(Name)} CW{-> MARK(Name, 1, 2, 3)};
CW{-PARTOF(Name), -REGEXP(".")} COMMA? FirstNameInitial+{-> MARK(Name, 1, 2, 3)};
```

Implementations of Entity Extraction


Rule-Based
Hybrid
Machine Learning Based
NLP Papers (2003-2012)
All Vendors
Large Vendors
Commercial Vendors (2013)