Relevance-based abstraction identification: technique and evaluation

R. Gacitua, P. Sawyer, and V. Gervasi.
Requir. Eng., 16(3):251-265, 2011.

Itzel Morales Ramirez
Fondazione Bruno Kessler
Trento, Italy

Outline

- Introduction
- Problem: discovery of key abstractions in domain docs
- Techniques: automatic term recognition (ATR)
- Evaluation methodology: precision and recall measures
- Discussion & Conclusion

Definitions

- Document is made of tokens (“words”)
- Terms are sequences of tokens
- Term is a potential signifier of an underlying abstraction
- Abstraction is a concept with particular significance
- Extraction is the derivation of a set of abstractions from a doc

Motivation

- Usefulness in requirements engineering (RE)
  - Quick understanding of the problem domain
  - Domain expertise contained in documents
  - Abstractions inferred from words or phrases
  - Several useful purposes: project dictionary, validation list, verification of requirements

Problem: discovery of key abstractions

- *D*: document
- *K:* static sources of knowledge, not dependent on D
- A_D*:* we want to extract a set of terms
- $\sigma : A_D \rightarrow R$, measure of significance
- $\leq_\sigma$ ranking, higher values are more significant

$$\forall a_1, a_2 \in A_D, a_1 \leq_\sigma a_2 \iff \sigma(a_1) \leq \sigma(a_2)$$

Quality of the set depends on 2 factors

1) Includes all and only significant terms
2) Degrees of significance (ranking)

Techniques: ATR

- Relevance-based Abstraction Identification (RAI)
  - Designed to identify terms
  - Use of statistical methods
    - Rank candidate abstractions
  - Techniques used
    - Corpus-based frequency profiling – individual words
      - LL_w: log-likelihood value to determine the significance of words

$$LL_w = 2\left(w_d \cdot \ln \frac{w_d}{E_d} + w_c \cdot \ln \frac{w_c}{E_c}\right)$$

Occurrence of word w in k, i.e. corpus, and document

    - Syntactic patterns - multiword
      - Posit multiword terms as common combinations of adjectives and nouns, adverbs and verbs, and prepositions

RAI-0

- Hypothesis: not all the words contribute equally to the significance value of multiword term of which they are a component

- Significance value for multiword terms
  - Term t = $\langle w_1, w_2 \ldots, w_l \rangle$
  - *k :* weight to each word based on its position

$$S_t = \frac{\sum_i k_i LL_{w_i}}{l}$$

Hypothesis based on Term is composed of a headword *sailing ship*

adjective modifier

noun headword

RAI-0 procedure

- Annotation of words with a Part-of-Speech (i.e. linguistic categories: noun, verb…)
- Filtering of words unlikely to signify abstractions
- Reduction of words to their lemma (e.g. run and running -> run)
- Assignment of LL value to each word, where *K*= British National Corpus
- Application of syntactic patterns
- Calculation of the significance score for every term (S_t)
- Sorting of terms based on significance value

Evaluation methodology

- Measure of success
  - Existence of a reference set of abstractions A'_D
  - Precision: how many of the abstractions we extracted were relevant

$$Precision = \frac{|A_D \cap A'_D|}{A_D}$$

  - Recall: how many of the relevant abstractions we could extract

$$Recall = \frac{|A_D \cap A'_D|}{A'_D}$$

  - Genuine abstractions: *lag* metric, defined as the average number of false positives (higher significance value)

$$lag = \frac{\sum_{a \in A'_D} lag(a)}{|A'_D|}$$

lower lag -> genuine abstractions

Case study

- D: book on a technical domain, i.e. RFID
- A'_D: analytical index

- D contains 156,028 words
- A'_D has 911 entries

- Simplifying assumption due to the absence of ranking
  - 911 abstractions are of equal importance
  - Recall and precision measures by counting the topmost 5, 10, …, 20 terms

Evaluation

- Term “RFID tag”
- *k*: constant weight 1.0 vs variable weight

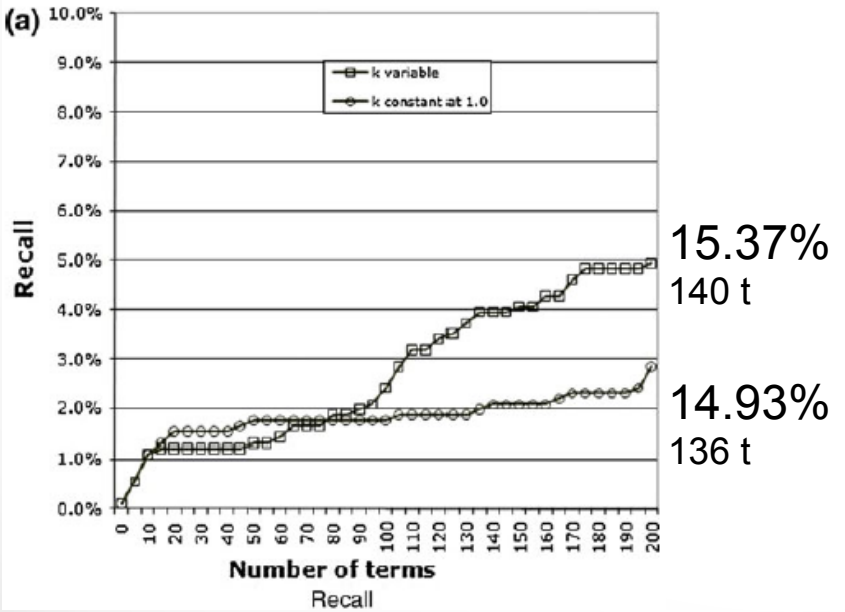
Hypothesis

$$t = \langle w_1, w_2 \ldots, w_l \rangle$$

$$S_t = \frac{\sum_i k_i LL_{w_i}}{l}$$

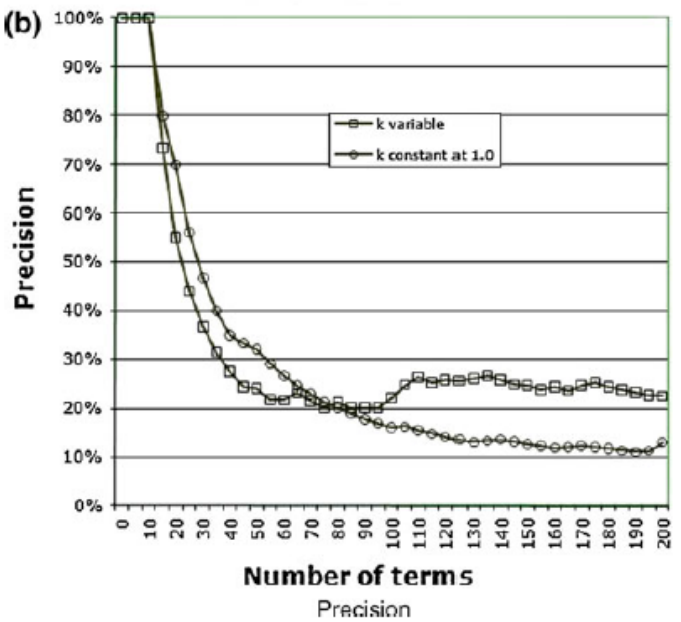
l=1 k_0=1.0
l=2 k_1=0.4, k_0=0.6
l=3 k_2=0.2, k_1=0.3, k_0=0.5

k weight
LL_{RFID}=0.4 LL_{tag}=0.6



15.37%
140 t

14.93%
136 t



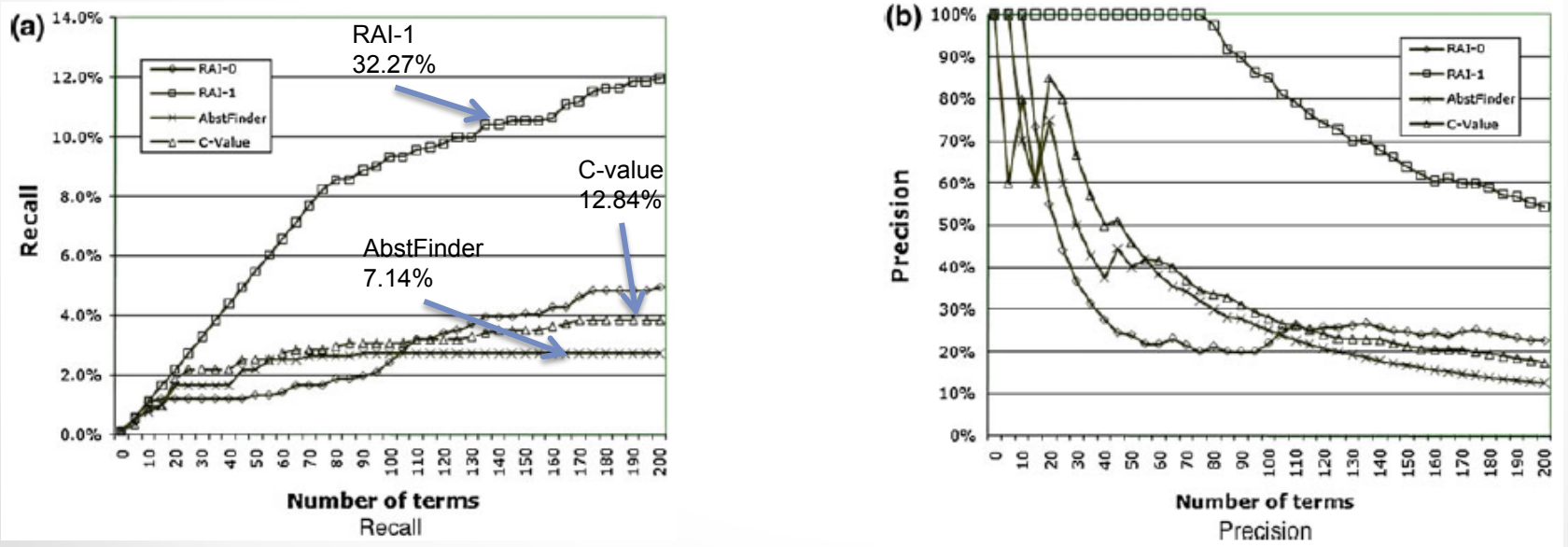
15.37%
14.93%

Improvements

- RAI-1
  - It might be defined a richer set of syntactic patterns
  - Solution: combination of corpus-based frequency with raw frequency

$$S'_t = S_t \frac{termFreq_D(t)}{\max_{t' \in A_D} termFreq_D(t')}$$

Comparison of RAI-1 vs RAI-0, AbstFinder and C-Value



Discussion

- Limitations
  - Book index is a tough test
  - Human judgement is needed
- Threads to validity
  - Lexical similarity does not guarantee semantic relatedness
  - The RFID book is a single, coherent text written by 1 author

Conclusion

- New technique for the identification of single- and multiword abstractions (RAI)
- The implemented tool provides guidance to understand the domain
- Investigation of feedback from requirements engineers to filter abstractions or extract further ones

Thank you!

Any questions?