Koninklijke Bibliotheek
National library of the Netherlands

Max Planck Institute for Psycholinguistics

vrije Universiteit amsterdam

NWO
Netherlands Organisation for Scientific Research

Putting ontology alignment in context:

Usage scenarios, deployment and evaluation in a library case

Antoine Isaac
Henk Matthezing
Lourens van der Meij
Stefan Schlobach
Shenghui Wang
Claus Zinn

TEL plus project

Introduction

- Alignment technology can help solving important problems
  - heterogeneity of description resources
- But:
  - *What for, exactly?*
  - *How useful can it be?*

- Consensus: generation and evaluation of alignment have to take into account **applications**
- Problem: (relatively) not much investigation on alignment applications and their requirements

Putting alignment into context: approach

- Focusing on application **scenarios**

  For a given scenario

  - What are the expected meaning and use of alignments?
  - How to use results of current alignment tools?
  - How to fit evaluation to application's success criteria?

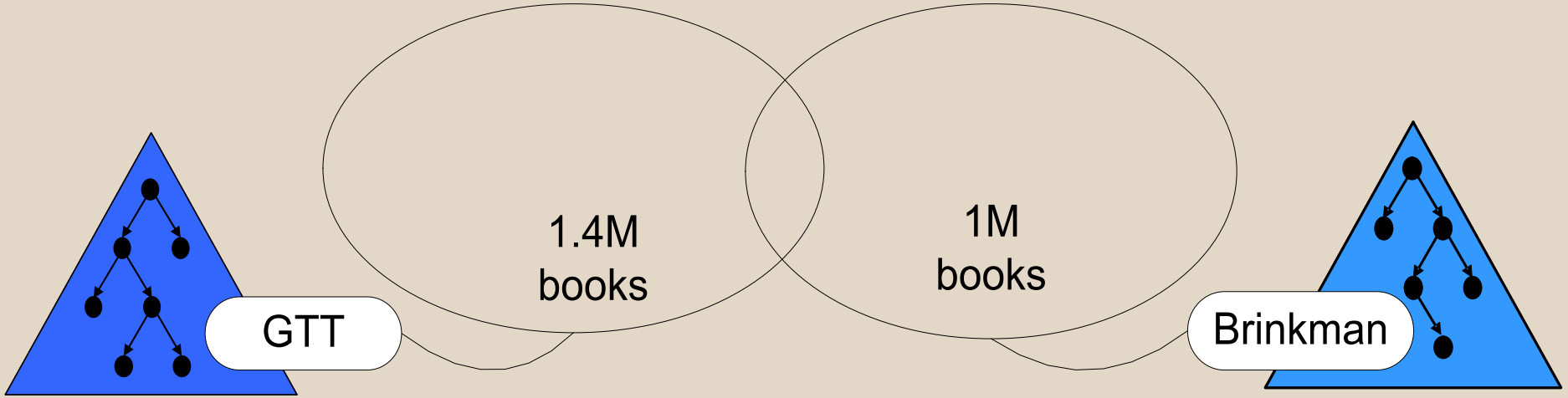
- Testing two hypotheses
  - For a same scenario, different evaluation strategies can bring different results
  - For two scenarios, evaluation results can differ for a same alignment, even with the most appropriate strategies

Agenda

- The KB application context
- Focus on two scenarios
  - Thesaurus merging
  - Book re-indexing
- OAEI 2007 Library track scenario-specific evaluation

Our application context

- National Library of the Netherlands (KB)
- 2 main collections
- Each described (*indexed*) by its own thesaurus



Usage scenarios for thesaurus alignment at KB

- Concept-based search
  - Retrieving GTT-indexed books using Brinkman concepts
- Book re-indexing
  - Indexing GTT-indexed books with Brinkman concepts
- Integration of one thesaurus into the other
  - Inserting GTT elements into the Brinkman thesaurus
- Thesaurus merging
  - Building a new thesaurus from GTT and Brinkman
- Free-text search
  - matching user search terms to *both* GTT or Brinkman concepts
- Navigation
  - browse the 2 collections through a merged version of the thesauri

Agenda

- The KB application context
- Focus on two scenarios
  - Thesaurus merging
  - Book re-indexing
- OAEI 2007 Library track scenario-specific evaluation

Thesaurus merging scenario

- Merge two vocabularies in a single, unified one
- Requirement: for two concepts, say whether a (thesaurus) semantic relation holds
  - Broader (BT), narrower (NT), related (RT)
  - Equivalence (EQ), if the two concepts share a same meaning and should be merged in a single one

- Similar to ontology engineering cases

  [Euzenat & Shvaiko, 2007]

Deploying alignments for thesaurus merging

- De facto standard for alignment results
  `(e1,e2,relation,measure)`

- Problem: relation
  - "=", `rdfs:subClassOf` or `owl:equivalentClass`
  - Adaption has to be made
    - Danger of overcommitment or loosening

- Problem: confidence/similarity measure
  - Meaning?
  - Weighted mappings have to be made crisp (e.g. by threshold)

Thesaurus merging: evaluation method

- Alignments are evaluated in terms of individual mappings
  - Does the mapping relation apply?
  - Quite similar to classical ontology alignment evaluation
- Mappings can be assessed directly

In a thesaurus combining these two concepts, which relationship would hold between them?

. **Adrenalectomie**	○ narrower than [4] ◉ equivalent to [5] ○ broader than [6] ○ related to [7] ○ no link [8] ○ don't know [9]	bedrijfsadministratie . **loonadministratie**

Thesaurus merging evaluation measures

- **Correctness**: proportion of proposed links that are correct
- **Completeness**: how many correct links were retrieved
- IR measures of **precision** and **recall** against a gold standard can be used
  - Eventually semantic versions [Euzenat]

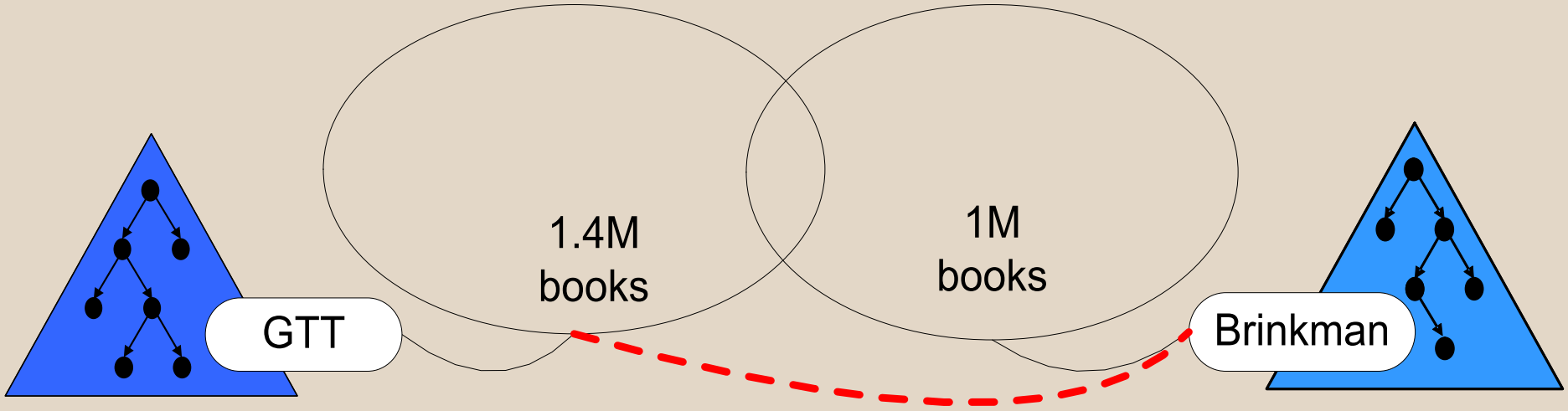
- Note: when no gold standard is present, other measures for completeness can be considered:
  - **coverage over a set of proposed alignments**, for *comparative* evaluation of alignment tools
  - **coverage over the thesauri** can be helpful for practitioners

Agenda

- The KB application context
- Focus on two scenarios
  - Thesaurus merging
  - Book re-indexing
- OAEI 2007 Library track scenario-specific evaluation

Book re-indexing scenario

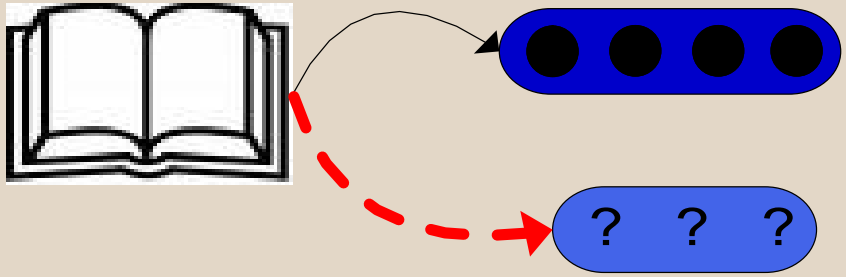
- Scenario: re-annotation of GTT-indexed books by Brinkman concepts



- If one thesaurus is dropped, legacy data has to be indexed according to the other voc.
  - Automatically or semi-automatically

Scientific Collection

Book re-indexing requirements



- Requirement for a re-indexing function: converting sets of concepts to sets of concepts

- **post-coordination**: co-occurrence matters
  `{G1="History" , G2="the Netherlands"}` for GTT
  a book about Dutch history
- **granularity** of two vocabularies differ
  `{B1="Netherlands; History"}` for Brinkman

Semantic interpretation of re-indexing function

One-to-one case: g1 can be converted to b1 if:

- Ideal case: b1 is semantically equivalent to g1
- But b1 could also be more general than g1
  - Loss of information
  - OK if b1 is the most specific subsumer of g1's meaning
    - Indexing specificity rule
- …

Deploying alignments for book re-indexing

- Results of existing tools may need re-interpretation
- Unclear semantics of mapping relations and weights
  - As for thesaurus merging
- Single concepts involved in mappings
  - We need conversion of *sets* of concepts
  - Only a few matching tools perform multi-concept mappings

[Euzenat & Shvaiko]

Deploying alignments for book re-indexing

- Solution: generate *rules* from 1-1 mappings

```
"Sport" exactMatch "Sport"
+ "Sport" exactMatch "Sport practice"
    => "Sport" -> {"Sport", "Sportpractice"}
```

  - Several aggregation strategies are possible

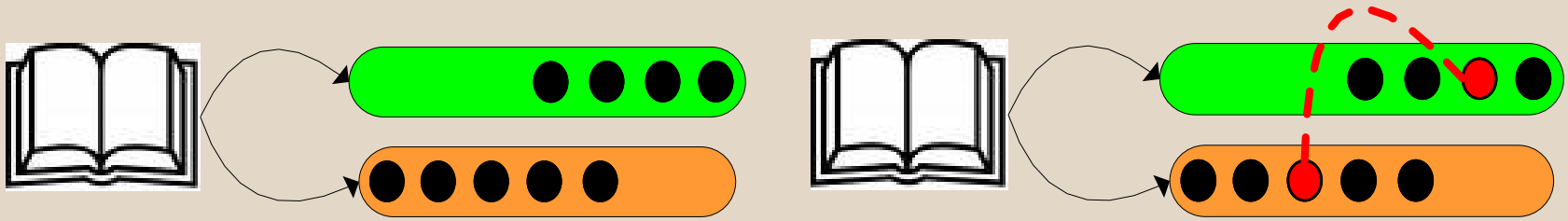
- Firing rules for books
  - Several strategies, e.g. fire a rule for a book if its index includes rule's antecedent

- Merge results to produce new annotations

Re-indexing evaluation

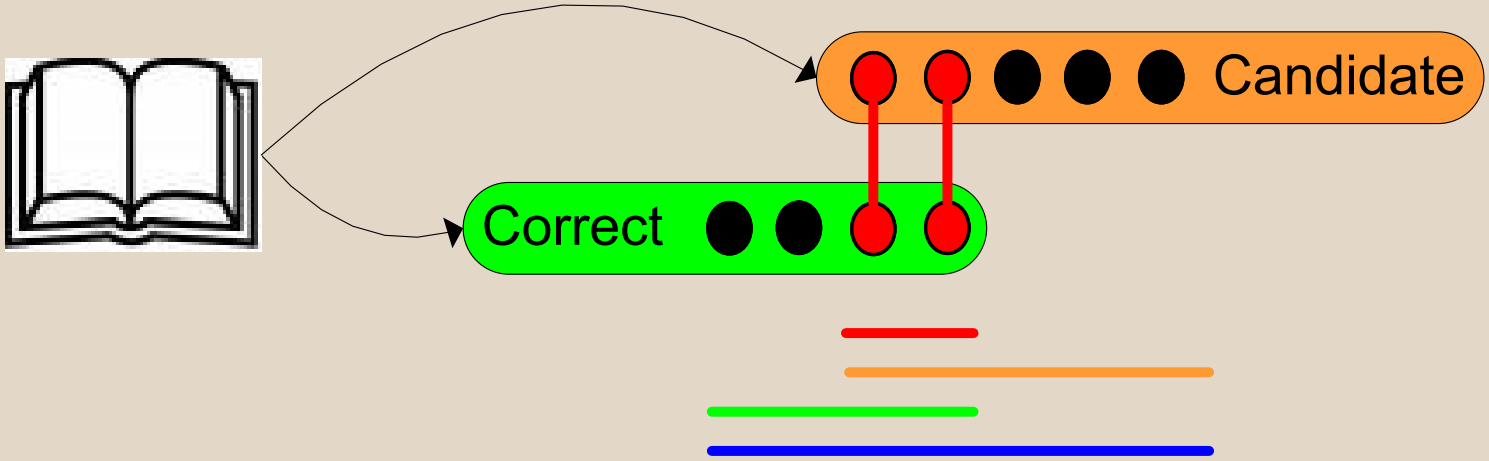
- We do not assess the mappings, nor even the rules
- We assess their application for book indexing
  - More *end-to-end*

- General method: compare the annotations produced with the alignment with the correct ones



Re-indexing evaluation measures

- Annotation level: measure correctness and completeness of the set of produced concepts
  - Precision, Recall, Jaccard overlap (general distance)



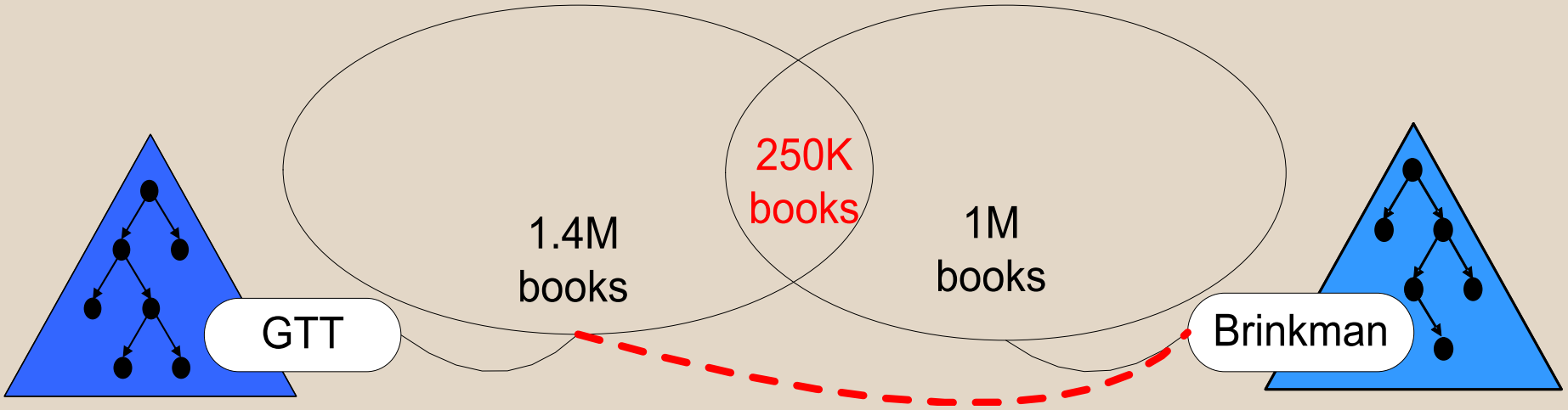
- Notice: counting over annotated books, not rules or concepts
  - *Rules and concepts used more often are more important*

Re-indexing evaluation measures

- Book level: counting *matched* books
  - Books for which there is one good annotation
  - Minimal hint about users' (dis)satisfaction

Re-indexing: automatic evaluation

- There is a gold standard!



Human evaluation vs. automatic evaluation

Problems when considering application constraints

- Indexing variability
  - Several indexers may make different choices
  - Automatic evaluation compares with a specific one
- Evaluation variability
  - Only one expert judgment is considered per book indexing assessment
- Evaluation set bias
  - Dually-indexed books may present specific characteristics

Re-indexing: manual evaluation

- Human expert assesses candidate indices: would have they chosen the same concepts?
  - A “maybe” answer is now possible
- A slightly different perspective on evaluation criteria
  - *Acceptability* of candidate indices

Agenda

- The KB application context
- Focus on two scenarios
  - Thesaurus merging
  - Book re-indexing
- OAEI 2007 Library track scenario-specific evaluation

Ontology Alignment Evaluation Initiative (OAEI)

- Apply and evaluate aligners on different tracks/cases
- Campaigns organized since 2004, and every year
  - More tracks, more realistic tracks
  - Better results of alignment tools

  Important for scientific community!

- OAEI 2007 Library track: KB thesauri
- Participants: Falcon, DSSim, Silas
  - Mostly exactMatch-mappings

http://oaei.inrialpes.fr/

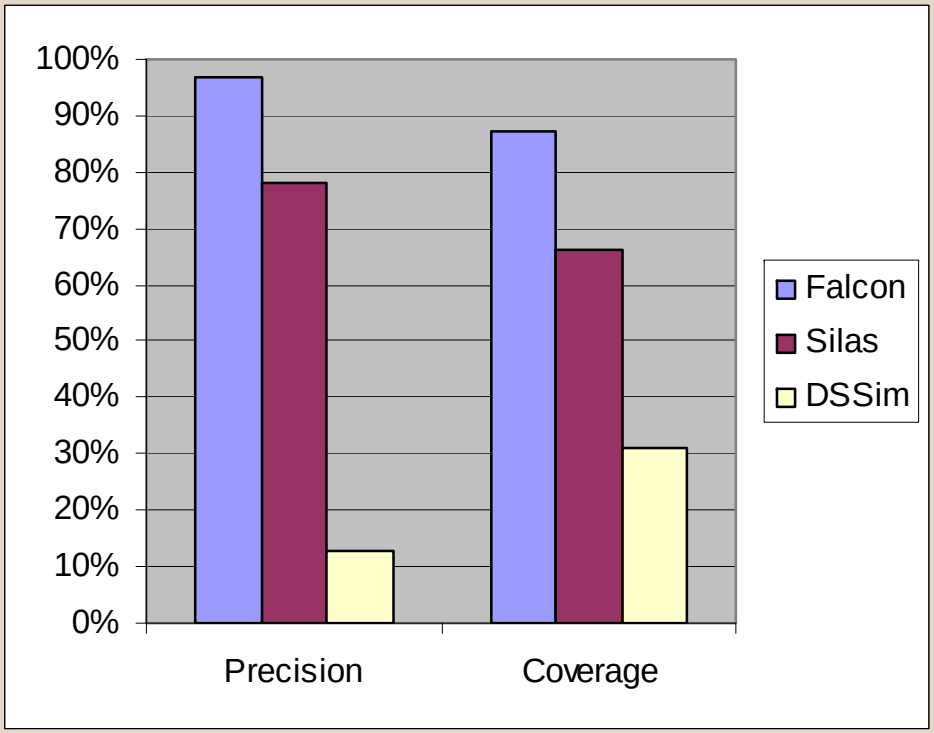
Thesaurus merging evaluation

- No gold standard available

- Comparison with “reference” lexical alignment
- Manual assessment for a sample of “extra” mappings

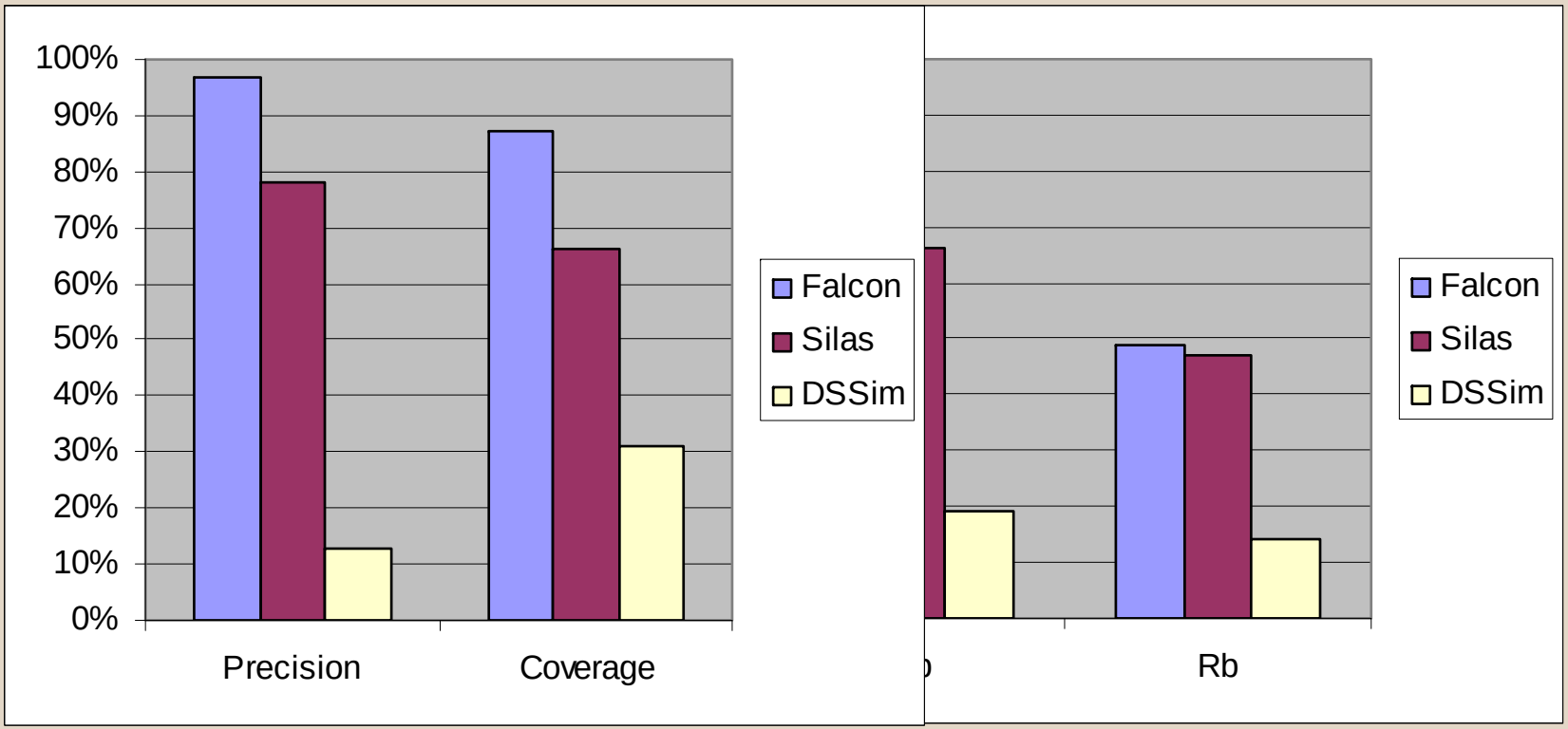
- *Coverage:* proportion of good mappings found (participants + reference)

Thesaurus merging: evaluation results



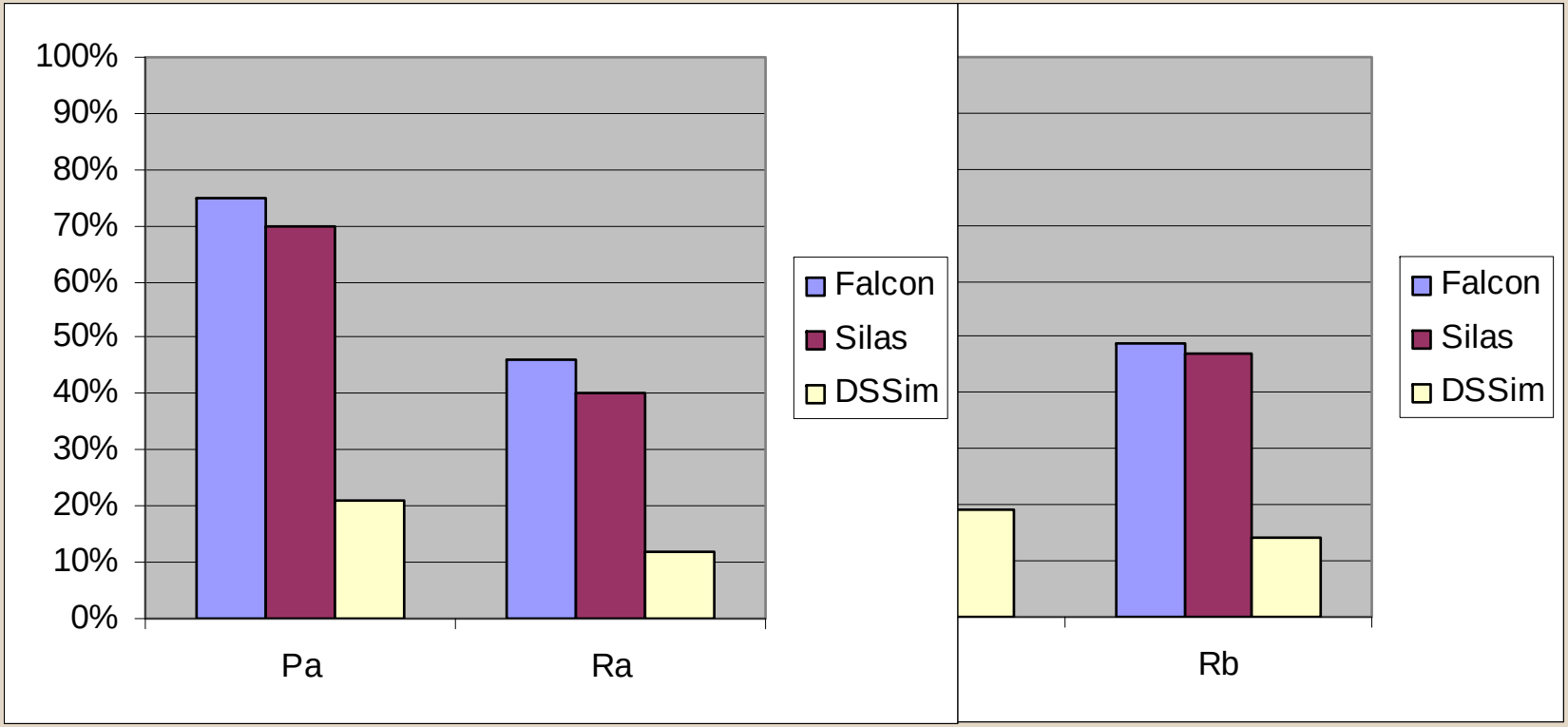
- Falcon performs well: closest to lexical reference
- DSSim and Ossewaarde add more to the lexical reference
  - Ossewaarde adds less than DSSim, but additions are better

Book re-indexing: automatic evaluation results



Book re-indexing: manual evaluation results

Research question: *quality of candidate annotations*



- *Performances are consistently higher than for automatic evaluation*

Book re-indexing: manual evaluation results

- Research question: *evaluation variability*
  - Jaccard overlap between evaluators' assessments: 60%
  - Krippendorff's agreement coefficient (alpha): 0.62
- Research question: *indexing variability*
  - For dually indexed books, almost 20% of original indices are not even acceptable!

Conclusions: observations

- Variety of scenarios requiring alignment
- There are common requirements, but also differences
- Leading to different success criteria and evaluation measures

- There is a gap with current alignment tools
  - Deployment efforts are required
- Confirmation that different alignment strategies perform differently on different scenarios
  - Choosing appropriate strategies

Take-home message

- Just highlighting flaws?
- No, application-specific evaluation also helps to improve state-of-the-art alignment technology

- Simple but necessary things
  - Evaluation is not free
  - When done carefully, it brings many benefits

Thanks!