Learning Concept Mappings from Instance Similarity

Shenghui Wang[1] Gwenn Englebienne[2] Stefan Schlobach[1]

[1] Vrije Universiteit Amsterdam

[2] Universiteit van Amsterdam

ISWC 2008

vrije Universiteit amsterdam

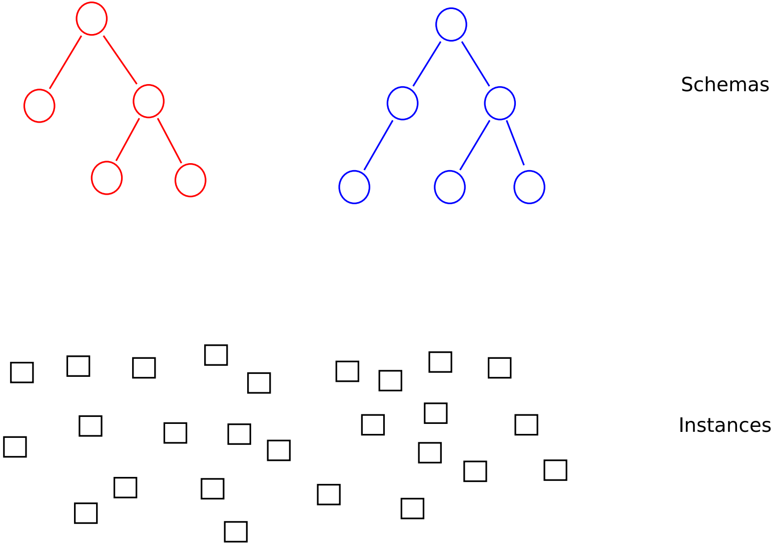
University of Amsterdam

Outline

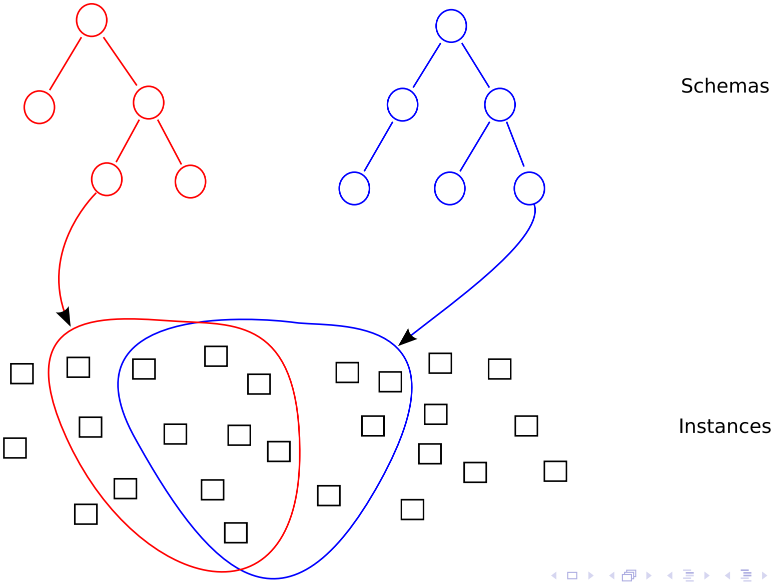
Thesaurus mapping

- SemanTic Interoperability To access Cultural Heritage (STITCH) through mappings between thesauri
  - *e.g.* "*plankzeilen*" (board sailing) *vs.* "*surfsport*" (surfing)
  - *e.g.* "*griep*" (flu) *vs.* "*influenza*"
- Scope of the problem:
  - Big thesauri with tens of thousands of concepts
  - Huge collections (*e.g.*, KB: 80km of books in one collection)
  - Heterogeneous (*e.g.*, books, manuscripts, illustrations, *etc.*)
  - Multi-lingual problem

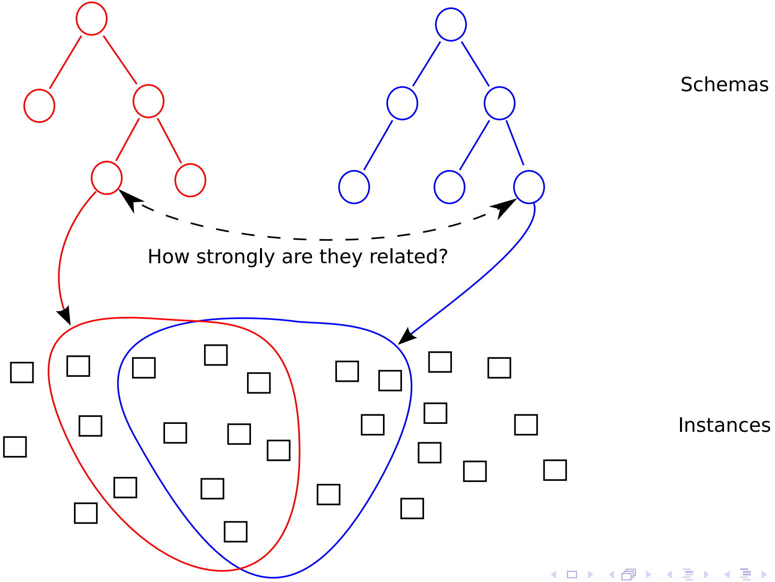
Instance-based techniques: common instance based



Instance-based techniques: common instance based



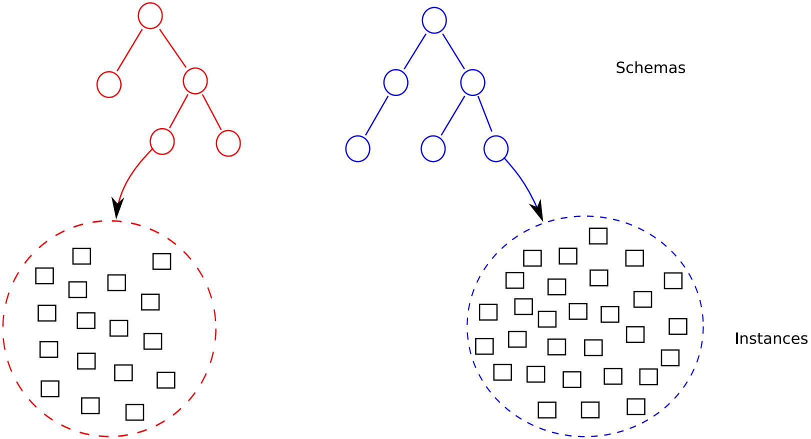
Instance-based techniques: common instance based



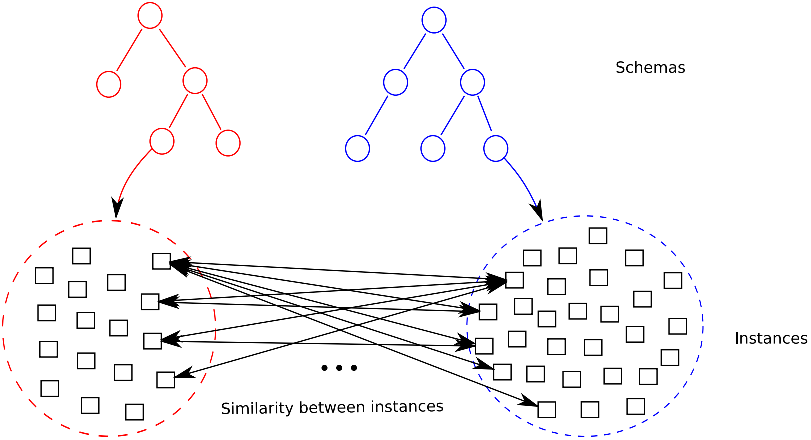
Pros and cons

- Advantages
  - Simple to implement
  - Interesting results
- Disadvantages
  - Requires sufficient amounts of common instances
  - Only uses part of the available information

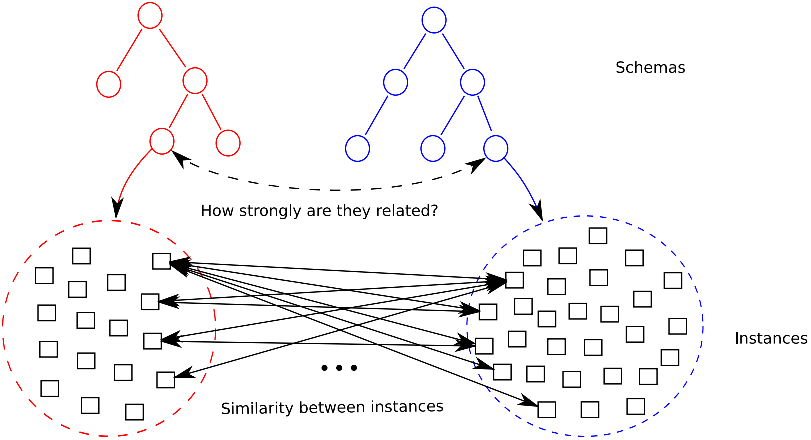
Instance-based techniques: Instance similarity based



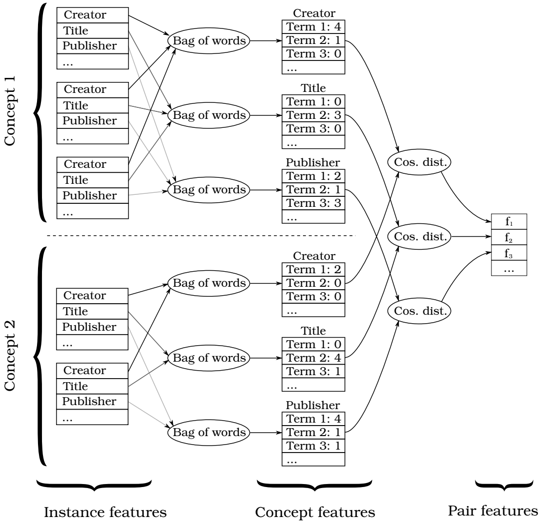
Instance-based techniques: Instance similarity based



Instance-based techniques: Instance similarity based



Representing concepts and the similarity between them



Classification based on instance similarity

- Each pair of concepts is treated as a point in a "similarity space"
  - Its position is defined by the features of the pair.
  - The features of the pair are the different measures of similarity between the concepts' instances.
- Hypothesis: the *label* of a point — which represents whether the pair is a *positive* mapping or *negative* one — is correlated with the position of this point in this space.
- With already labelled points and the actual similarity values of concepts involved, it is possible to classify a point, *i.e.*, to give it a right label, based on its location given by the actual similarity values.

The classifier used: Markov Random Field

- Let $T = \{ (\mathbf{x}^{(i)}, y^{(i)}) \}_{i=1}^{N}$ be the training set
  - $\mathbf{x}^{(i)} \in \mathbb{R}^K$, the features
  - $y^{(i)} \in \mathcal{Y} = \{positive, negative\}$, the label
- The conditional probability of a label given the input is modelled as

$$p(y^{(i)}|\mathbf{x}_i, \theta) = \frac{1}{Z(\mathbf{x}_i, \theta)} \exp\left(\sum_{j=1}^{K} \lambda_j \phi_j(y^{(i)}, \mathbf{x}^{(i)})\right), \qquad (1)$$

where $\theta = \{ \lambda_j \}_{j=1}^{K}$ are the weights associated to the feature functions ϕ and $Z(\mathbf{x}_i, \theta)$ is a normalisation constant

The classifier used: Markov Random Field (cont')

- The likelihood of the data set for given model parameters $p(T|\theta)$ is given by:

$$p(T|\theta) = \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)}) \tag{2}$$

  During learning, our objective is to find the most likely values for θ for the given training data.

- The decision criterion for assigning a label $y^{(i)}$ to a new pair of concepts i is then simply given by:

$$y^{(i)} = \operatorname*{argmax}_{y} p(y|\mathbf{x}^{(i)}) \tag{3}$$

Research questions

- Are the benefits from feature-similarity of instances in extensional mapping significant?
- **Joint or non-joint** Can our approach be applied to corpora for which there are no dually annotated instances?
- **Heterogeneous collections** Can our approach be applied to corpora in which instances are described in a heterogeneous way?
- **Feature weighting** Can we make qualitative use of the learned model?

Experiment setup

- Two cases:
  - mapping GTT (35K) and Brinkman (5K) used in Koninklijke Bibliotheek (KB) — Homogeneous collections
  - mapping GTT/Brinkman and GTAA (160K) used in Beeld en Geluid (BG) — Heterogeneous collections
- Evaluation
  - Measures: misclassification rate or error rate
  - 10 fold cross-validation
  - testing on special data sets

Experiment I: Feature-similarity based mapping versus existing methods

Are the benefits from feature-similarity of instances in extensional mapping significant when compared to existing methods? Yes

Mapping method	Error rate
Falcon	0.28895
S_{lex}	0.42620 ± 0.049685
S_{jacc80}	0.44643 ± 0.059524
S_{bag}	0.57380 ± 0.049685
$\{f_1, \ldots f_{28}\}$ **(our new approach)**	$\mathbf{0.20491 \pm 0.026158}$

Table: Comparison between existing methods and similarities-based mapping, in KB case

Experiment I: Feature-similarity based mapping versus existing methods

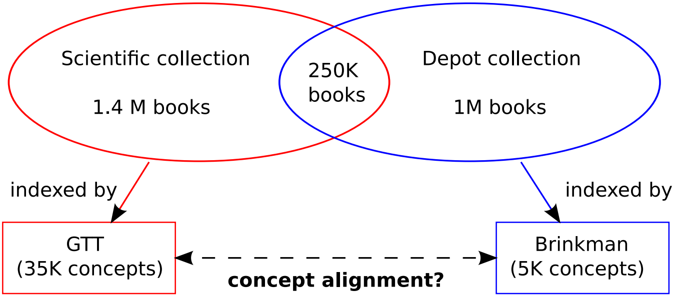
Are the benefits from feature-similarity of instances in extensional mapping significant when compared to existing methods? **Yes**

Mapping method	Error rate
Falcon	0.28895
S_{lex}	0.42620 ± 0.049685
S_{jacc80}	0.44643 ± 0.059524
S_{bag}	0.57380 ± 0.049685
$\{f_1, \ldots f_{28}\}$ **(our new approach)**	$\mathbf{0.20491 \pm 0.026158}$

Table: Comparison between existing methods and similarities-based mapping, in KB case

Experiment II: Extending to corpora without joint instances

Can our approach be applied to corpora for which there are no doubly annotated instances, *i.e.*, for which there are no joint instances?



Experiment II: Extending to corpora without joint instances (cont')

Yes

Collections	Testing set	Error rate
Joint instances (original KB corpus)	golden standard	0.20491 ± 0.026158
	lexical only	0.137871
No joint instances (double instances removed)	golden standard	0.28378 ± 0.026265
	lexical only	0.161867

Table: Comparison between classifiers using joint and disjoint instances, in KB case

Experiment III: Extending to heterogeneous collections

Can our approach be applied to corpora in which instances are described in a heterogeneous way?

- Feature selection
  - exhaustive combination by calculating the similarity between all possible pairs of fields
    - require more training data to avoid over-fitting
  - manual selection of corresponding metadata field pairs
  - mutual information to select the most informative field pairs

Experiment III: Extending to heterogeneous collections

Can our approach be applied to corpora in which instances are described in a heterogeneous way?

- Feature selection
  - exhaustive combination by calculating the similarity between all possible pairs of fields
    - require more training data to avoid over-fitting
  - manual selection of corresponding metadata field pairs
  - mutual information to select the most informative field pairs

Experiment III: Extending to heterogeneous collections

Can our approach be applied to corpora in which instances are described in a heterogeneous way?

- Feature selection
  - exhaustive combination by calculating the similarity between all possible pairs of fields
    - require more training data to avoid over-fitting
  - manual selection of corresponding metadata field pairs
  - mutual information to select the most informative field pairs

Experiment III: Extending to heterogeneous collections

Can our approach be applied to corpora in which instances are described in a heterogeneous way?

- Feature selection
  - exhaustive combination by calculating the similarity between all possible pairs of fields
    - require more training data to avoid over-fitting
  - manual selection of corresponding metadata field pairs
  - mutual information to select the most informative field pairs

Experiment III: Extending to heterogeneous collections

Can our approach be applied to corpora in which instances are described in a heterogeneous way?

- Feature selection
  - exhaustive combination by calculating the similarity between all possible pairs of fields
    - require more training data to avoid over-fitting
  - manual selection of corresponding metadata field pairs
  - mutual information to select the most informative field pairs

Feature selection

Can we maintain high mapping quality when features are selected (semi)-automatically?

Yes

Thesaurus	Feature selection	Error rate
GTAA vs. Brinkman	manual selection	0.11290 ± 0.025217
	mutual information	$\mathbf{0.09355 \pm 0.044204}$
	exhaustive	0.10323 ± 0.031533
GTAA vs. GTT	manual selection	0.10000 ± 0.050413
	mutual information	$\mathbf{0.07826 \pm 0.044904}$
	exhaustive	0.11304 ± 0.046738

Table: Comparison of the performance with different methods of feature selection, using non-lexical dataset

Feature selection

Can we maintain high mapping quality when features are selected (semi)-automatically?
Yes

Thesaurus	Feature selection	Error rate
GTAA vs. Brinkman	manual selection	0.11290 ± 0.025217
	mutual information	$\mathbf{0.09355} \pm \mathbf{0.044204}$
	exhaustive	0.10323 ± 0.031533
GTAA vs. GTT	manual selection	0.10000 ± 0.050413
	mutual information	$\mathbf{0.07826} \pm \mathbf{0.044904}$
	exhaustive	0.11304 ± 0.046738

Table: Comparison of the performance with different methods of feature selection, using non-lexical dataset

Training set

- manually built golden standard (751)
- lexical seeding
- background seeding

Thesauri	lexical	non-lexical
GTAA *vs.* GTT	2720	116
GTAA *vs.* Brinkman	1372	323

Table: Numbers of positive examples in the training sets

Training set

- manually built golden standard (751)
- lexical seeding
- background seeding

Thesauri	lexical	non-lexical
GTAA *vs.* GTT	2720	116
GTAA *vs.* Brinkman	1372	323

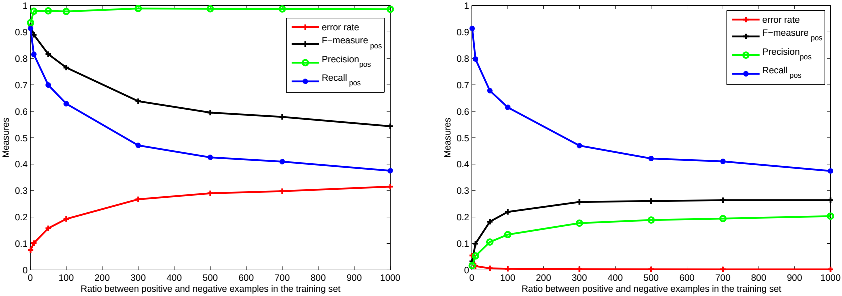
Table: Numbers of positive examples in the training sets

Bias in the training sets

Thesauri	Training set	Test set	Error rate
GTAA *vs.* Brinkman	non-lexical	non-lexical	0.09355 ± 0.044204
	lexical	non-lexical	0.11501
	non-lexical	lexical	0.07124
	lexical	lexical	0.04871 ± 0.029911

Table: Comparison using different datasets (feature selected using mutual information)

Positive-negative ratios in the training sets



(a) testing on 1:1 data (b) testing on 1:1000 data

Figure: The influence of positive-negative ratios — Brinkman *vs.* GTAA

Positive-negative ratios in the training sets (cont')

In practice, the training data should be chosen so as to contain a **representative ratio** of positive and negative examples, while still providing enough material for the classifier to have good **predictive capacity** on both types of examples.

Experiment IV: Meta-data mapping

The value of learning results, λ_j, reflects the importance of the feature f_j in the process of determining similarity (mappings) between concepts.

KB fields	BG fields
kb:title	bg:subject
kb:abstract	bg:subject
kb:annotation	bg:LOCATIES
kb:annotation	bg:SUBSIDIE
kb:creator	bg:contributor
kb:creator	bg:PERSOONSNAMEN
kb:Date	bg:OPNAMEDATUM
kb:dateCopyrighted	bg:date
kb:description	bg:subject
kb:publisher	bg:NAMEN
kb:temporal	bg:date

Experiment IV: Meta-data mapping

The value of learning results, λ_j, reflects the importance of the feature f_j in the process of determining similarity (mappings) between concepts.

KB fields	BG fields
kb:title	bg:subject
kb:abstract	bg:subject
kb:annotation	bg:LOCATIES
kb:annotation	bg:SUBSIDIE
kb:creator	bg:contributor
kb:creator	bg:PERSOONSNAMEN
kb:Date	bg:OPNAMEDATUM
kb:dateCopyrighted	bg:date
kb:description	bg:subject
kb:publisher	bg:NAMEN
kb:temporal	bg:date

Summary

- We use a machine learning method to automatically use the similarity between instances to determine mappings between concepts from different thesauri/ontologies.
  - Enables mappings between thesauri used for very heterogeneous collections
  - Does not require dually annotated instances
  - Not limited by the language barrier
  - A contribution to the field of meta-data mapping
- In the future
  - More heterogeneous collections
  - *Smarter* measures of similarity between instance metadata
  - More similarity dimensions between concepts, *e.g.*, lexical, structural

Thank you

Computation complexity

- Training: based on an iterative Quasi-Newton method (LBFGS) which is quite efficient but iterative, depending on where you started and how precise you want your answer to be
- Inference: linear in the number of features