Entity Deduplication on ScholarlyData

Ziqi Zhang, Nottingham Trent University, UK
Andrea Giovanni Nuzzolese, STLab, ISTC-CNR, Italy
Anna Lisa Gentile, IBM Research, US

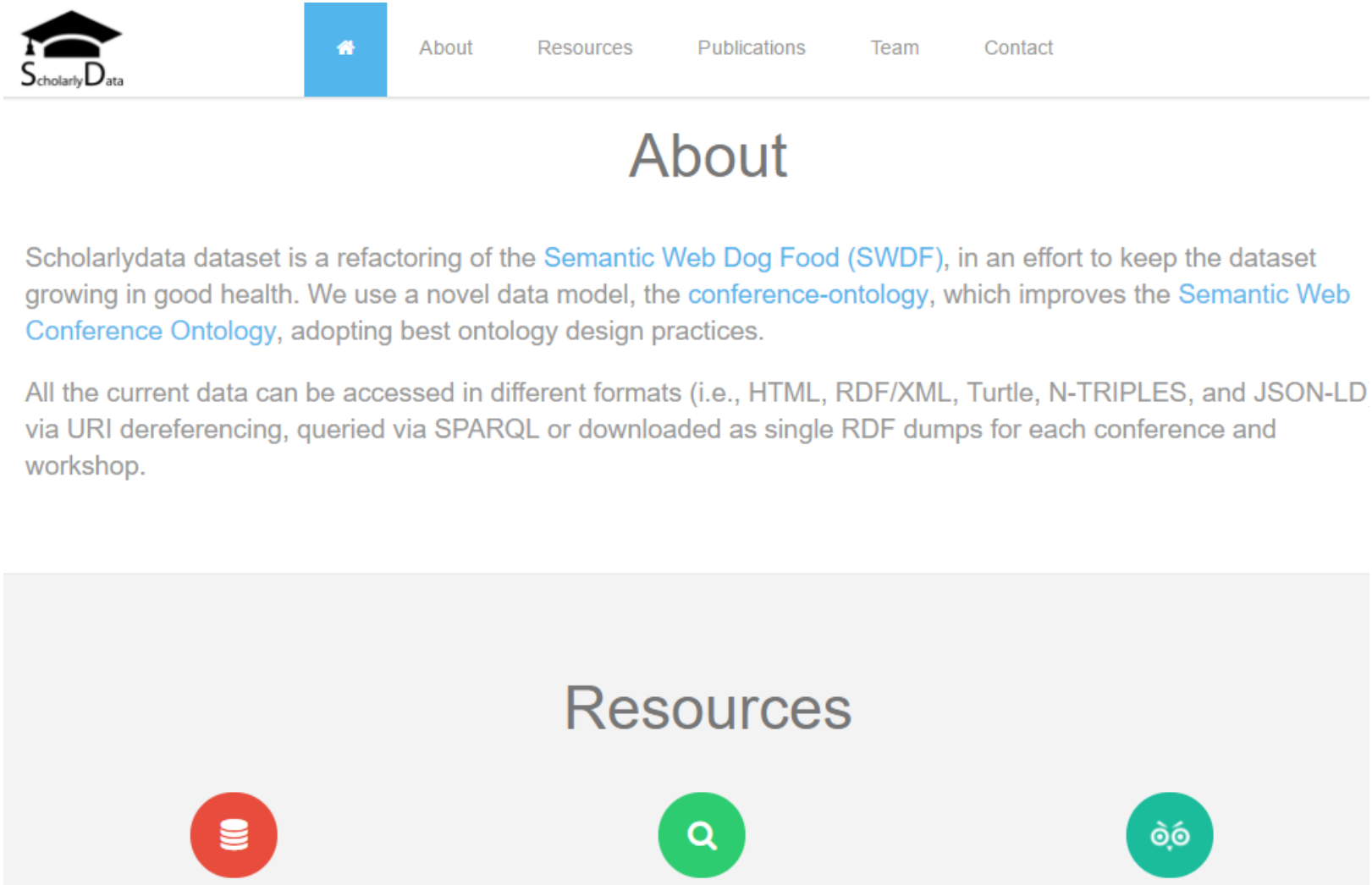
Outline

- BACKGROUND
  - ScholarlyData
  - Related Work: conference linked data, URI/entity de-duplication, duplicate URI harmonization
- SUPERVISED MACHINE LEARNING
  - Blocking
  - Classification
  - URI harmonization
- EVALUATION
- CONCLUSION AND FUTURE WORK

BACKGROUND

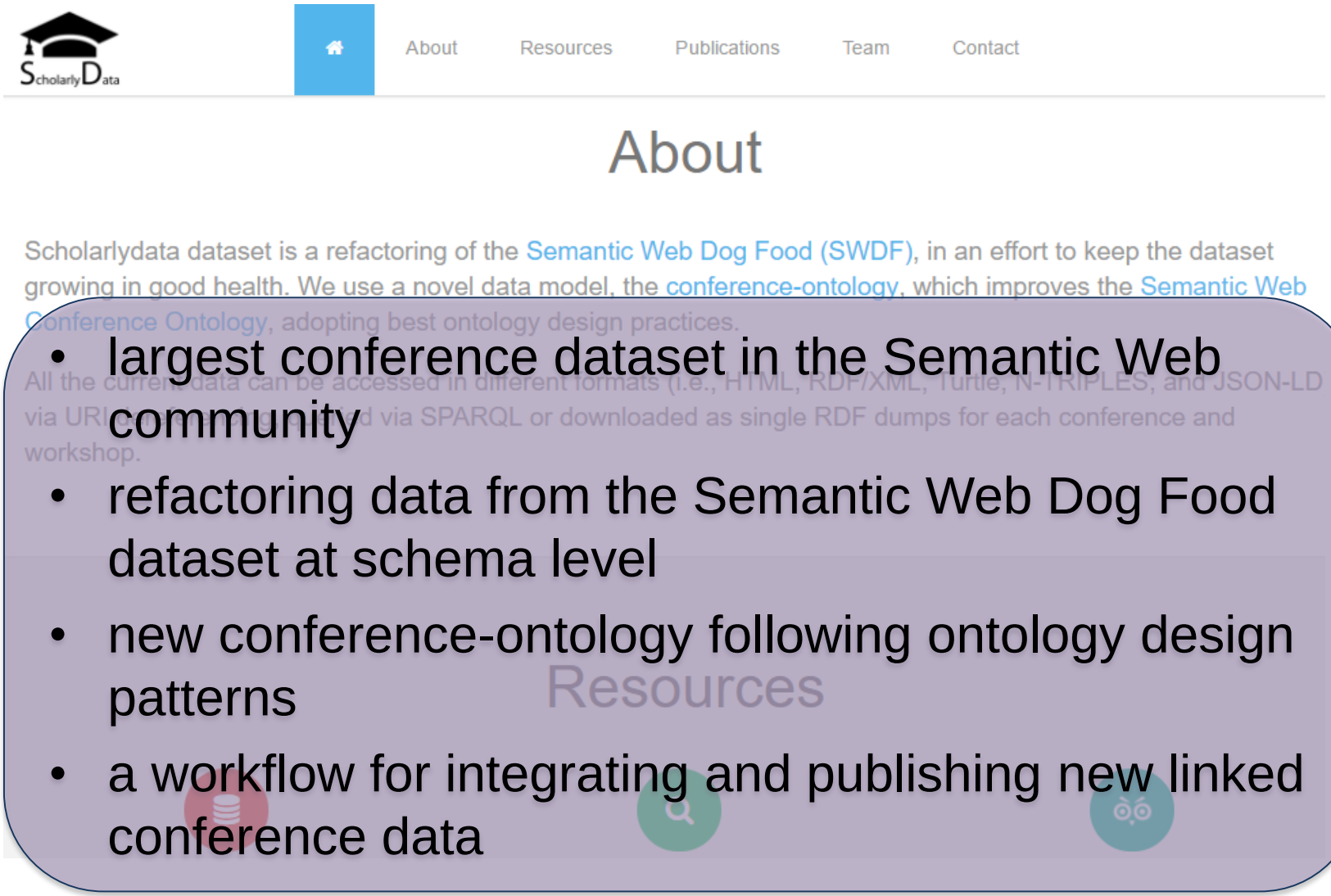
BACKGROUND

ScholarlyData.org – what is it?



BACKGROUND

ScholarlyData.org – what is it?



- largest conference dataset in the Semantic Web community
- refactoring data from the Semantic Web Dog Food dataset at schema level
- new conference-ontology following ontology design patterns
- a workflow for integrating and publishing new linked conference data

BACKGROUND

ScholarlyData.org – problems

About https://w3id.org/scholarlydata/organisation/free-university-of-berlin

Property: http://www.w3.org/2002/07/owl#sameAs

http://data.semanticweb.org/organization/free-university-of-berlin

Subject: https://w3id.org/scholarlydata/affiliation-during-event/ekaw2012-freie-univ
paschke

About https://w3id.org/scholarlydata/organisation/freie-universitaet-berlin

Property: http://www.w3.org/2002/07/owl#sameAs

http://data.semanticweb.org/organization/freie-universitaet-berlin

- Duplicates at **instance** level
  - historical data issues
  - practice of URI re-use not followed by publishers
- Most prominent entity types: PERSON, ORGANIZATION

BACKGROUND

Related Work

- Conference linked datasets
  - Started with metadata project ESWC2006, ISWC2006
  - ‘Semantification’ of conference datasets promoted by major publishers and data indexers
  - **ScholarlyData – largest conference linked dataset and leading development in this area**
- URI/entity de-duplication
  - Link discovery [Nentwig et al. 2015] & OAEI ontology matching – instance matching [Euzenat et al. 2013]
  - **Novelty: domain adaptation + features to combat unbalanced usage of duplicate (co-referent) URIs**

BACKGROUND

Related Work

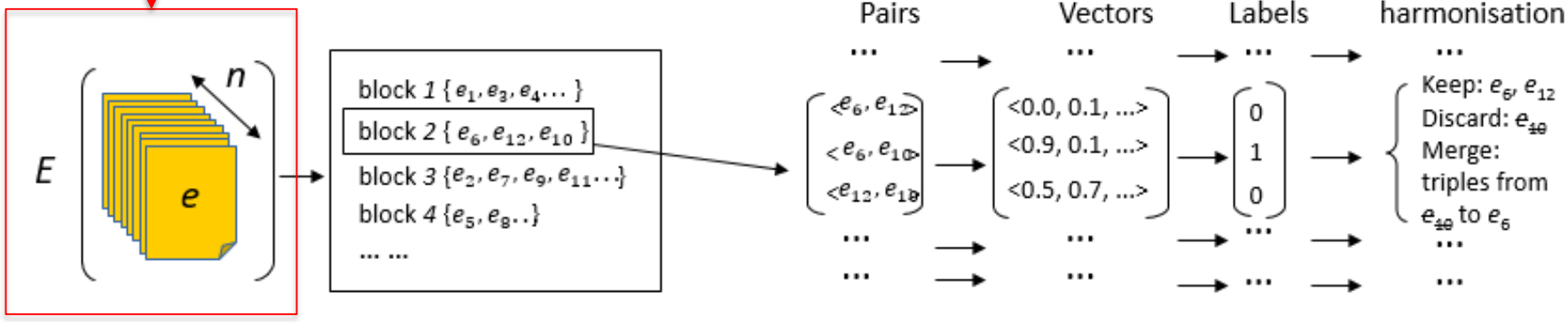
- Co-referent URI harmonization
  - owl:SameAs not suitable for co-reference/duplicate URIs [Glaser et al. 2009]
  - **Solution: HTTP redirect for dereferencing, SPARQL query re-writing for backward compatibility**

METHODOLOGY

Overview

METHODOLOGY

Input:
A set of URIs (E) of the same type, e.g., PERSON

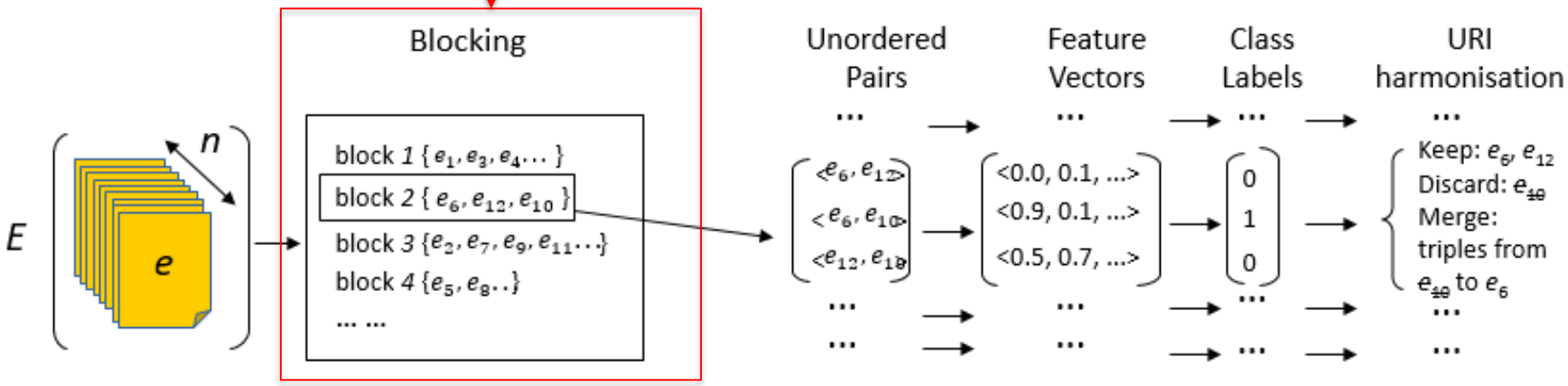


Overview

METHODOLOGY

Blocking:
Light-weight process to split URIs into groups of potentially similar ones for pair-wise comparison

- reduces the number of quadratic pair-wise comparison to $m \ll \binom{n}{2}$

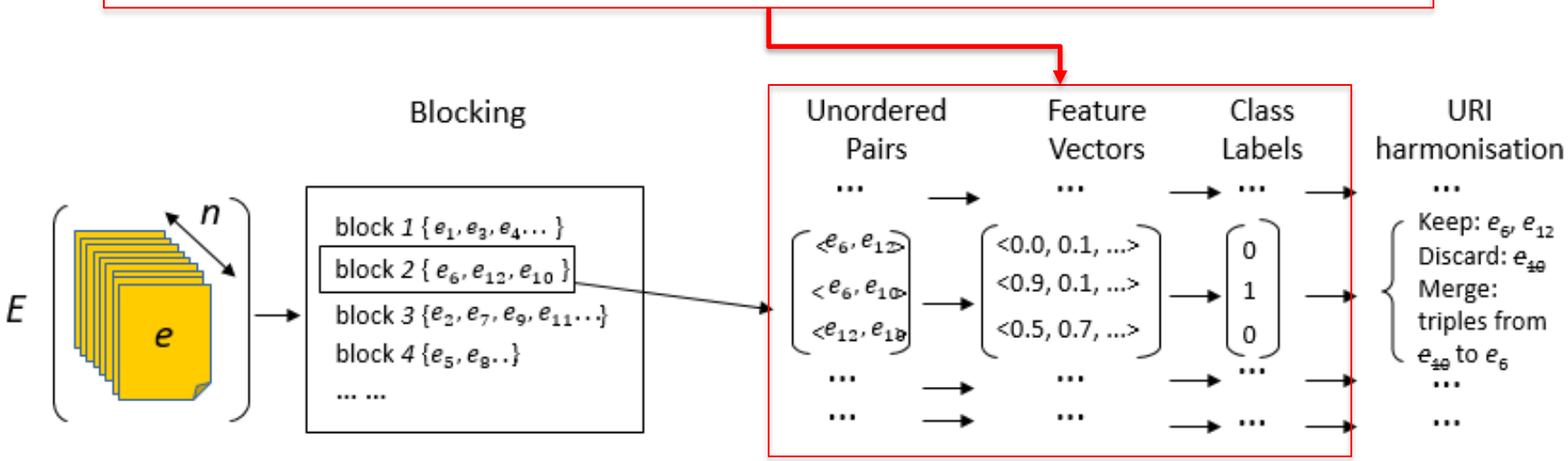


METHODOLOGY

Overview

Classification:
Given pairs of candidate URIs returned by **blocking**, determine for each pair if they are truly co-referent

- feature representation of URI pairs
- supervised classification

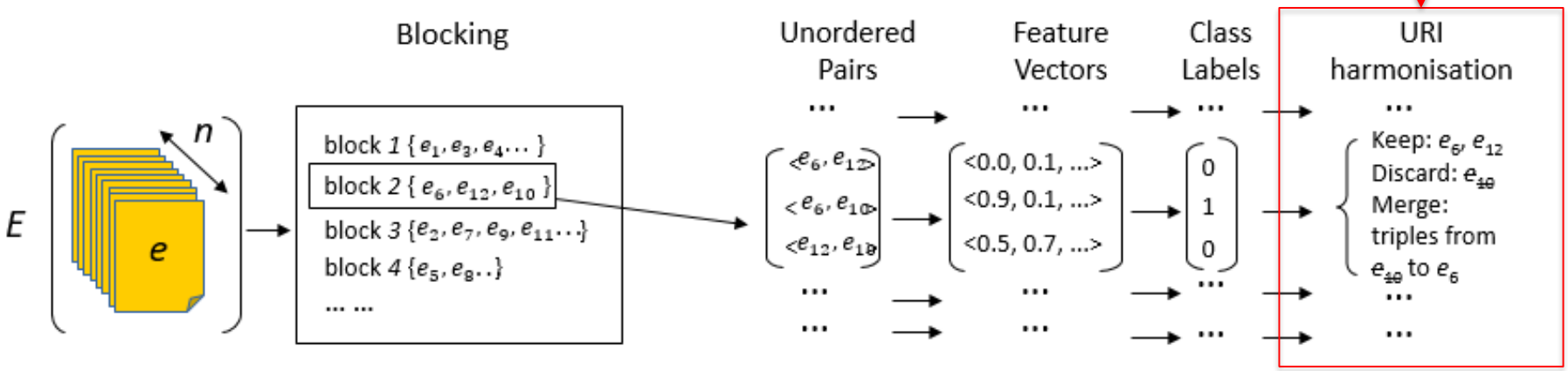


METHODOLOGY

Overview

URI harmonization:
Given co-referent URI pairs returned by **classification**, determine

- which URIs to keep and deprecate
- integrate data ensuring backward compatibility



Blocking

- Sorted neighborhood method (SNM)
  - sort all URIs by lexicographic order:
    - the URI string
    - literal value given the URI's predicate conf[1]:name
  - group sorted list of URIs by a fixed window
- Content based method
  - each URI grouped with all other URIs that share at least one common token in their name-like property values [Papadakis et al. 2011]

1. https://w3id.org/scholarlydata/ontology/conference-ontology.owl#

Classification

- Feature representation for each **URI** – traversal of linked data graph following pre-defined paths to create a 'bag-of-values' representation
  - **PERSON** (8 features, see paper for details):

Feature	Path	Target(s)	Example
name	⟨conf:name⟩	Literals	'Tom Mitchell', 'T. Mitchell'
affiliation names	⟨conf:hasAffiliation, conf:withOrganisation, conf:name⟩	Literals	'STLab-CNR', 'ISTC-CNR'
affiliation URIs	⟨conf:hasAffiliation, conf:withOrganisation⟩	URIs	sdo:cnr-istc-italy
participated event URIs	⟨conf:hasAffiliation, conf:during⟩	URIs	sde:ESWC2009/ eswc/2009
published work URIs	⟨conf:hasContent^{-1}, conf:hasItem^{-1}, conf:hasAuthorList^{-1}⟩	URIs	sdi:ekaw2012/ paper/demos/109
co-author URIs	⟨conf:hasContent^{-1}, conf:hasItem^{-1}, conf:hasItem, conf:hasContent⟩	URIs	sdp:aldo-gangemi, sdp:valentina-presutti
title + abstract + keywords	starting from each each 'published work URI': ⟨conf:title⟩ ∨ ⟨conf:abstract⟩ ∨ ⟨conf:keyword⟩	Literals	'This paper describes a ...'

Classification

- Feature representation for each **URI** – traversal of linked data graph following pre-defined paths to create a 'bag-of-values' representation
  - **ORGANIZATION** (4 features, see paper for details):

Feature	Path	Target(s)	Example
name	⟨conf:name⟩	Literals	'STLab-CNR', 'ISTC-CNR'
members' names	⟨conf:withOrganisation^{-1}, conf:isAffliationOf, conf:name⟩	Literals	'L. Page', 'S. Brin'
members' URIs	⟨conf:withOrganisation^{-1}, conf:isAffliationOf⟩	URIs	sdp:aldo-gangemi
participated event URIs	⟨conf:withOrganisation^{-1}, conf:during⟩	URIs	sde:ESWC2009/ eswc/2009

Classification cont.

- Feature representation for each **URI pair**
  - type-wise feature similarity

$$dice(e_i, e_j, t) = \frac{|f(e_i,t) \cap f(e_j,t)|}{|f(e_i,t) \cup f(e_j,t)|} \quad (1)$$

$$dice^{sqrt}(e_i, e_j, t) = \sqrt{sim_{dice}(f(e_i,t), f(e_j,t))} \quad (2)$$

$$cov(e_i, e_j, t) = \max\{\frac{|f(e_i,t) \cap f(e_j,t)|}{|f(e_i,t)|}, \frac{|f(e_i,t) \cap f(e_j,t)|}{|f(e_j,t)|}\} \quad (3)$$

$$cov^{sqrt}(e_i, e_j, t) = \sqrt{cov(e_i, e_j, t)} \quad (4)$$

  - a URI pair is then represented as a vector of the type-wise feature similarities computed by all functions above, i.e:
    - PER pair: 8x4=32 features; ORG pair: 4x4=16 features

Classification cont.

- Given a feature vector representing a URI pair, determine if they are co-referent
  - Models:
    - Stochastic Gradient Descent (SGD)
    - Logistic Regression (LR)
    - Random Forest (RF)
    - SVM linear kernel (SVM-l)
    - SVM radial basis function (SVM-rbf)

METHODOLOGY

URI harmonization

- Given a pair of co-referent URI determine which one to keep
  - Closure identification
  - Candidate selection
  - Dataset update
  - Recording harmonization

EVALUATION

EVALUATION

Dataset

- Manually annotated URI pairs of PERSON and ORGANIZATION using the ScholarlyData dataset

Type	Total pairs	Positive	Negative
PER	698	148	550
ORG	424	188	236

EVALUATION

Blocking

- Metrics
  - pair completeness (PC), i.e. the fraction of true positive pairs retained
  - reduction ratio (RR), i.e. the percentage of pairs discarded
  - harmonic mean (HM): tradeoff between PC and RR.

SNM name	RR	PC	HM	SNM URI	RR	PC	HM	Content based	RR	PC	HM
5	≈ 1	0.73	0.84	5	≈ 1	0.6	0.75	L	≈ 1	0.38	0.55
10	≈ 1	0.8	0.89	10	≈ 1	0.71	0.83	N	≈ 1	0.38	0.55
20	≈ 1	0.91	0.95	20	≈ 1	0.8	0.89	S	≈ 1	1	≈ 1
30	0.99	0.91	0.95	30	0.99	0.84	0.91	L+N+S	≈ 1	1	≈ 1
50	0.99	0.93	0.96	50	0.99	0.84	0.91				
70	0.99	0.93	0.96	70	0.99	0.85	0.91				
90	0.98	0.95	0.97	90	0.98	0.89	0.93				

N stands for `conf:name`, S for `conf:familyName`, L for `rdfs:label`

PERSON

Blocking cont.

- Metrics
  - pair completeness (PC), i.e. the fraction of true positive pairs retained
  - reduction ratio (RR), i.e. the percentage of pairs discarded
  - harmonic mean (HM): tradeoff between PC and RR.

SNM name	RR	PC	HM	SNM URI	RR	PC	HM	Content based	RR	PC	HM
5	≈ 1	1	≈ 1	5	≈ 1	0.49	0.66	P	≈ 1	0.84	0.91
10	0.99	1	≈ 1	10	0.99	0.59	0.74	N	≈ 1	1	≈ 1
20	0.99	1	0.99	20	0.99	0.65	0.78	L	≈ 1	1	≈ 1
30	0.98	1	0.99	30	0.98	0.66	0.79	P+N+L	≈ 1	1	≈ 1
50	0.97	1	0.99	50	0.97	0.68	0.8				
70	0.96	1	0.98	70	0.96	0.68	0.8				
90	0.95	1	0.98	90	0.95	0.71	0.81				

N stands for `conf:name`, P for affiliated person, L for `rdfs:label`

ORGANIZATION

Classification

- Training and testing:
  - 75% for training 25% for testing
  - Parameter tuning by grid search with 10-fold validation on the training set
- Five models: Stochastic Gradient Descent (SGD), Logistic Regression (LR), Random Forest (RF), SVM linear kernel (SVM-l), SVM radial basis function (SVM-rbf)
- State-of-the-art: SILK [Isele et al. 2012], LIMES [Ngomo et al. 2011] wombat simple (l-ws), wombat complete (l-wc)

Classification

EVALUATION

ORG		SGD	LR	RF	SVM-l	SVM-rbf
Positive examples	P	0.85	0.84	0.86	0.83	0.83
	R	0.8	0.8	0.82	0.86	0.86
	F1	0.83	0.82	0.84	0.85	0.85
Negative examples	P	0.83	0.82	0.84	0.87	0.87
	R	0.87	0.85	0.87	0.84	0.84
	F1	0.85	0.84	0.86	0.85	0.85
Total	P	0.84	0.83	0.85	0.85	0.85
	R	0.84	0.83	0.85	0.85	0.85
	F1	0.84	0.83	0.85	0.85	0.85

PER		SGD	LR	RF	SVM-l	SVM-rbf
Positive examples	P	0.78	0.88	0.92	0.77	0.88
	R	0.51	0.63	0.66	0.49	0.63
	F1	0.62	0.73	0.77	0.6	0.73
Negative examples	P	0.89	0.91	0.92	0.88	0.91
	R	0.96	0.98	0.99	0.96	0.98
	F1	0.93	0.94	0.95	0.92	0.94
Total	P	0.87	0.91	0.92	0.88	0.91
	R	0.87	0.91	0.92	0.87	0.91
	F1	0.86	0.90	0.91	0.86	0.90

Results of the five different models

ORG		RF	L-ws	L-wc	SILK
Positive examples	P	0.86	0.63	0.73	0.59
	R	0.82	0.64	0.49	1.0
	F1	0.84	0.64	0.59	0.74
Negative examples	P	0.84	0.72	0.69	1.0
	R	0.87	0.71	0.86	0.47
	F1	0.86	0.72	0.77	0.64
Total	P	0.85	0.68	0.70	0.70
	R	0.85	0.68	0.70	0.70
	F1	0.85	0.68	0.70	0.70

PER		RF	L-ws	L-wc	SILK
Positive examples	P	0.92	0.34	0.78	0.48
	R	0.66	0.63	0.17	0.80
	F1	0.77	0.44	0.28	0.6
Negative examples	P	0.92	0.85	0.79	0.94
	R	0.99	0.61	0.98	0.71
	F1	0.95	0.71	0.88	0.81
Total	P	0.92	0.62	0.79	0.79
	R	0.92	0.62	0.79	0.73
	F1	0.91	0.62	0.79	0.76

Results against SoA

EVALUATION

URI Harmonization

- The best performing blocking (content based) strategy and classification model (random forest) is applied to the entire ScholarlyData dataset
- All predicted pairs manually evaluated
  - 0.86 precision for PER, 0.7 precision for ORG
  - 94 pairs of PER and 531 pairs of ORG integrated

CONCLUSION & FUTURE WORK

Conclusion and Future Work

- Compared to SoA:
  - Notable improvement in both tasks on F1 over SoA
  - Better balance between precision and recall
  - Better performance on detecting positive pairs (which is more useful in practice)
- Future work:
  - Cope with sparse features of ‘under-used’ URIs
  - Implicit connections of features
  - Extension for general purpose link discovery

Thank You & Questions

References

- [1] M. Nentwig, M. Hartung, A.-C. N. Ngomo, and E. Rahm. A survey of current link discovery frameworks. Semantic Web, (Preprint):1-18, 2015.
- [2] Jérôme Euzenat and Pavel Shvaiko. Ontology Matching. Springer-Verlag, Heidelberg (DE), 2nd edition, 2013.
- [3] H. Glaser, A. Ja ri, and I. Millard. Managing co-reference on the semantic web. In Linked Data on the Web (LDOW2009), 2009
- [4] G. Papadakis, C. Niederee, and P. Fankhauser. Ecient Entity Resolution for Large Heterogeneous Information Spaces. pages 535-544, 2011.
- [5] R. Isele and C. Bizer. Learning expressive linkage rules using genetic programming. Proc. VLDB Endow., 5(11):1638-1649, July 2012.
- [6] Sherif, M., Ngomo, A., and Lehmann, J. 2017. WOMBAT: A Generalization Approach for Automatic Link Discovery. In proceedings of 17th ESWC