Statistical Knowledge Patterns: Identifying Synonymous Relations in Large Linked Datasets

Ziqi Zhang[1] Anna Lisa Gentile[1] Eva Blomqvist [2]
Isabelle Augenstein[1] Fabio Ciravegna[1]

(1) Department of Computer Science, The University of Sheffield, UK

(2) Department of Computer and Information Science, Linköping University, Sweden

Scope

- Linked Data - diverse vocabulary
  - what if we don't know the vocabulary
  - possible overlap in vocabularies

Scope

- Linked Data - diverse vocabulary
  - what if we don't know the vocabulary
  - possible overlap in vocabularies

http://dbpedia.org/resource/9%C2%BD_Weeks	9
http://dbpedia.org/resource/A.D._(miniseries)	"A.D."@en
http://dbpedia.org/resource/ATL_(film)	"ATL"@en
http://dbpedia.org/resource/A_Christmas_Carol_(1999_film)	"A Christmas Carol"@en
http://dbpedia.org/resource/A_Genius,_Two_Partners_and_a_Dupe	"A Genius, Two Partners and a Dupe"@en
http://dbpedia.org/resource/A_Girl_Cut_in_Two	"A Girl Cut in Two"@en
http://dbpedia.org/resource/A_Hole_in_the_Head	"A Hole in the Head"@en
http://dbpedia.org/resource/A_Midsummer_Night's_Dream_(1999_film)	"A Midsummer Night's Dream"@en
http://dbpedia.org/resource/	
http://dbpedia.org/resource/	
http://dbpedia.org/resource/	
http://dbpedia.org/resource/	
http://dbpedia.org/resource/	
http://dbpedia.org/resource/	
http://dbpedia.org/resource/Absolute_Power_(film)	"Absolute Power"@en
http://dbpedia.org/resource/Accattone	"Accattone"@en
http://dbpedia.org/resource/Addicted_to_Love_(film)	"Addicted to Love"@en
http://dbpedia.org/resource/Adventure_(1945_film)	"Adventure"@en
http://dbpedia.org/resource/After.Life	"After.Life"@en
http://dbpedia.org/resource/After_Hours_(film)	"After Hours"@en
http://dbpedia.org/resource/After_the_Wedding	"After the Wedding"@en
http://dbpedia.org/resource/Agatha_(film)	"Agatha"@en
http://dbpedia.org/resource/Air_America_(film)	"Air America"@en

```
select distinct ?film ?title where
{?film a <http://dbpedia.org/ontology/Film>.
?film <http://dbpedia.org/property/name> ?title.
} LIMIT 100
```

Scope

- Linked Data - diverse vocabulary
  - what if we don't know the vocabulary
  - possible overlap in vocabularies

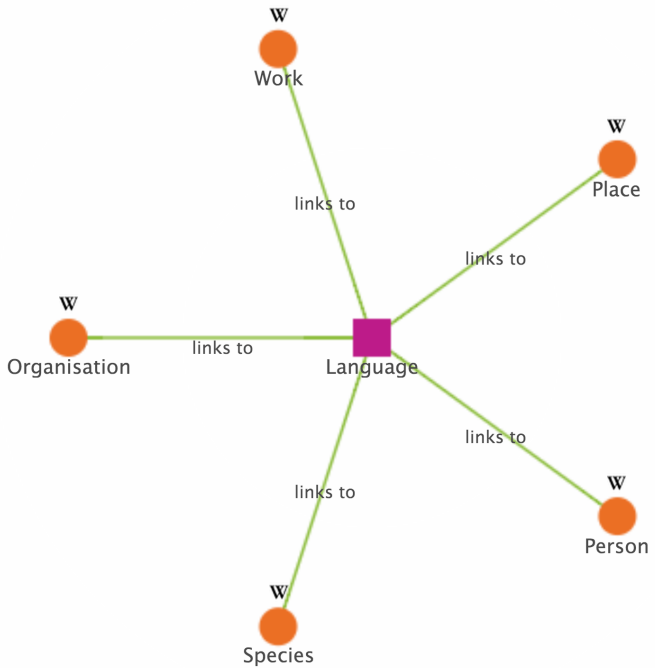
```
select distinct ?film ?title where
{?film a <http://dbpedia.org/ontology/Film>.
```

http://dbpedia.org/resource/9%C2%BD_Weeks	"9½ Weeks"@en
http://dbpedia.org/resource/A.D._(miniseries)	"A.D."@en
http://dbpedia.org/resource/ATL_(film)	"ATL"@en
http://dbpedia.org/resource/A_Christmas_Carol_(1999_film)	"A Christmas Carol"@en
http://dbpedia.org/resource/A_Genius,_Two_Partners_and_a_Dupe	"A Genius, Two Partners and a Dupe"@en
http://dbpedia.org/resource/A_Genius,_Two_Partners_and_a_Dupe	"(Un genio, due compari, un pollo)"@en
http://dbpedia.org/resource/A_Girl_Cut_in_Two	"A Girl Cut in Two"@en
http://dbpedia.org/resource/A_Hole_in_the_Head	"A Hole in the Head"@en
http://dbpedia.org/resource/A_Midsummer_Night's_Dream_(1999_film)	"A Midsummer Night's Dream"@en
http://dbpedia.org/resource/A_Nightmare_on_Elm_Street_(2010_film)	"A Nightmare on Elm Street"@en
http://dbpedia.org/resource/Absence_of_Malice	"Absence of Malice"@en
http://dbpedia.org/resource/Absolute_Power_(film)	"Absolute Power"@en
http://dbpedia.org/resource/Accattone	"Accattone"@en

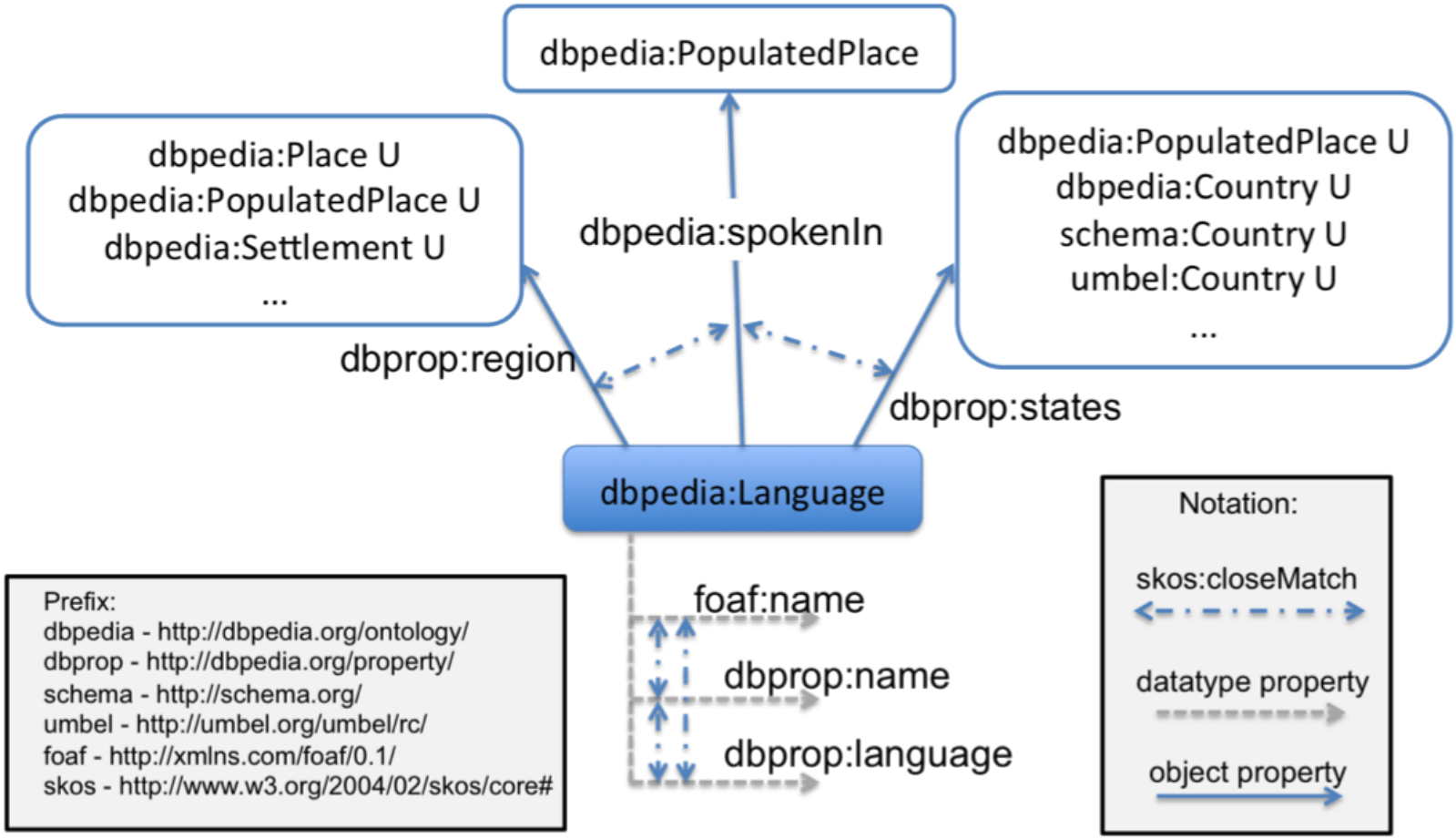
```
select distinct ?film ?title where
{?film a <http://dbpedia.org/ontology/Film>.
?film foaf:name ?title.
} LIMIT 100
```

Related efforts

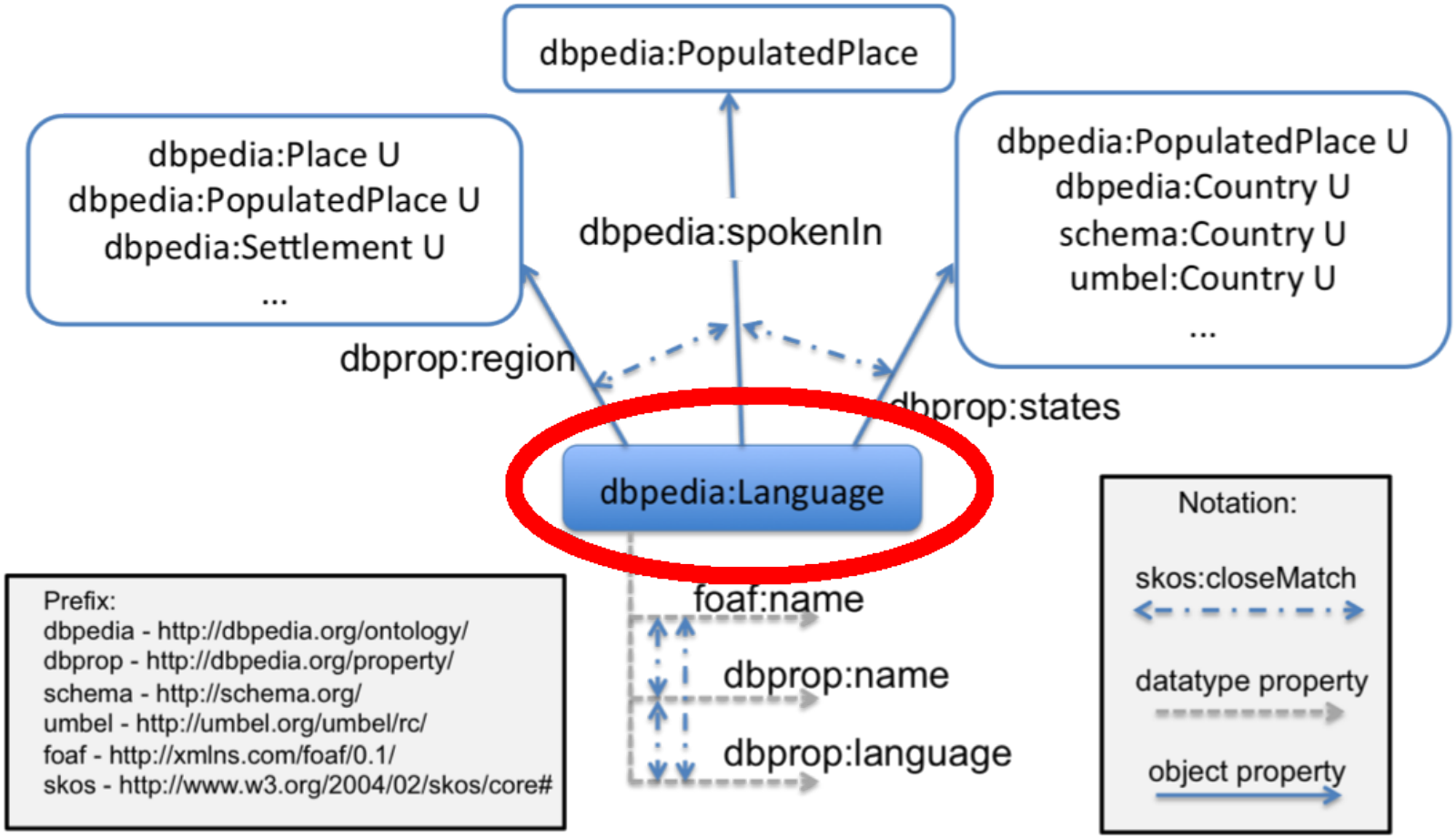
- *Encyclopedic Knowledge Patterns* (EKPs) [Nuzzolese et al., 2011]
  - Focused on data visualization and filtering, rather than querying
  - based on wikilinks rather than data representation
- Statistical characterization of datasets



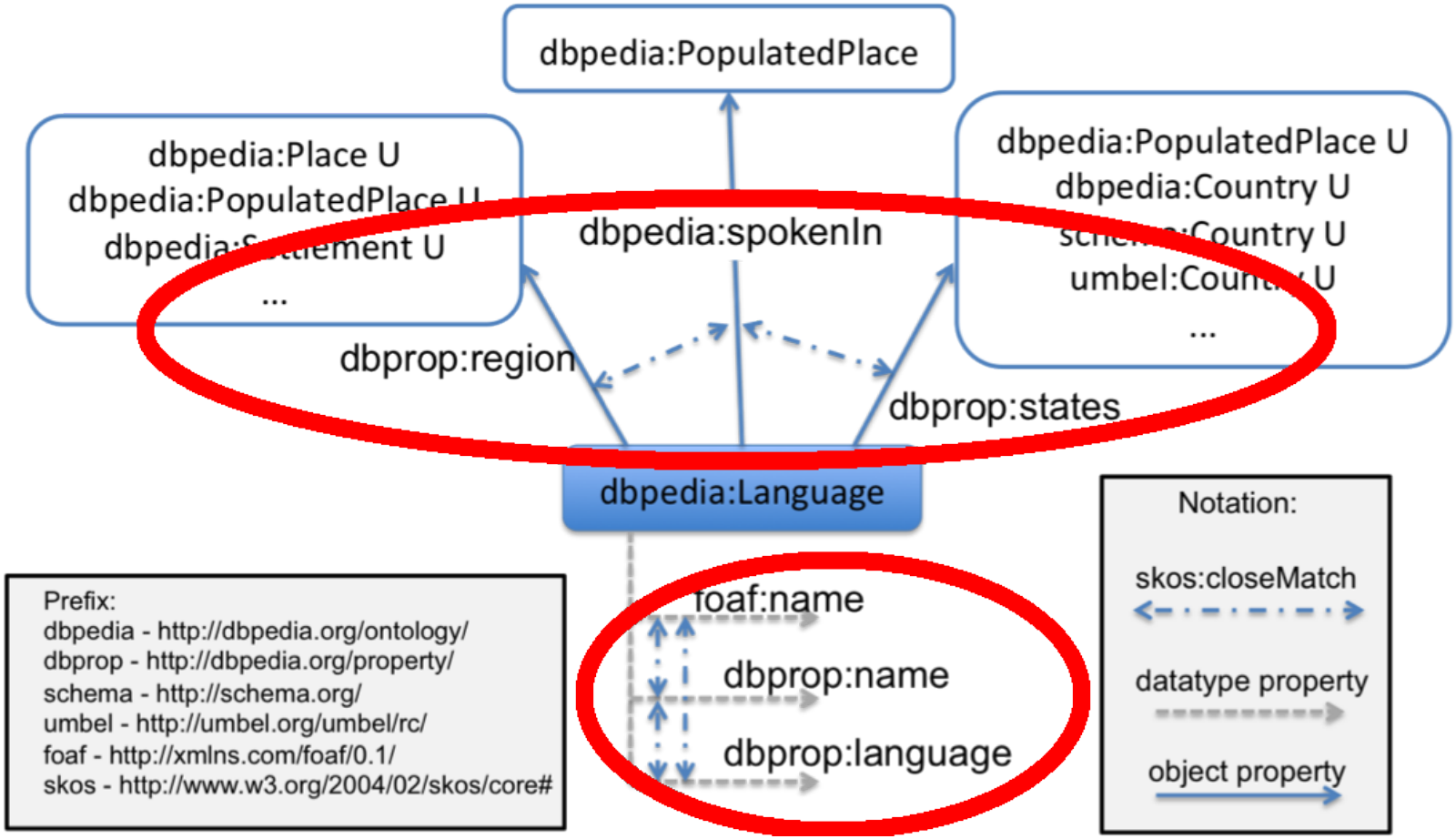
SKP Construction Overview



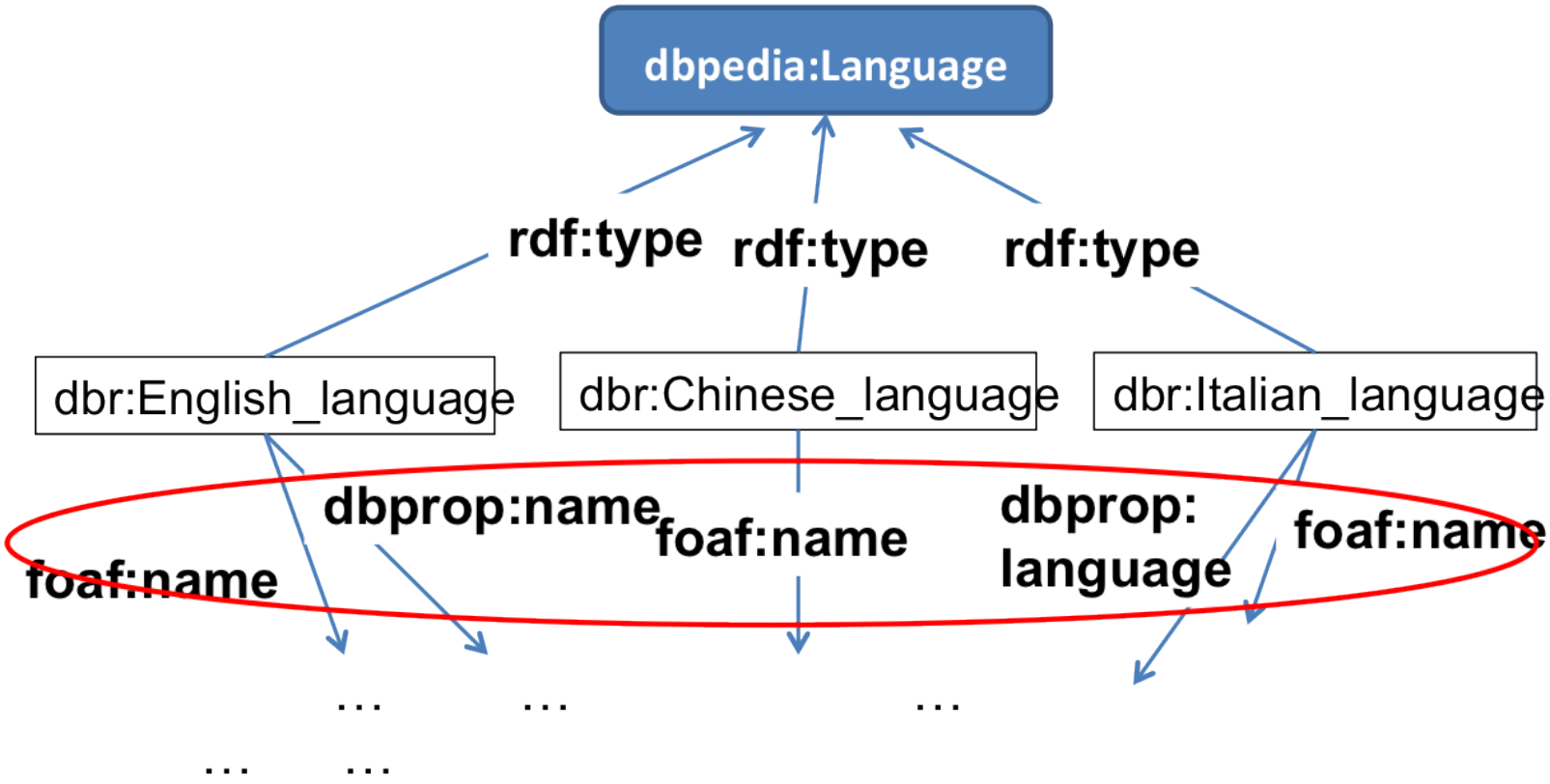
SKP Construction Overview



SKP Construction Overview

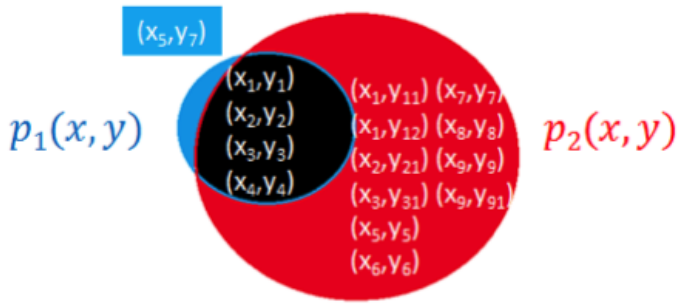


Finding Synonymous Properties



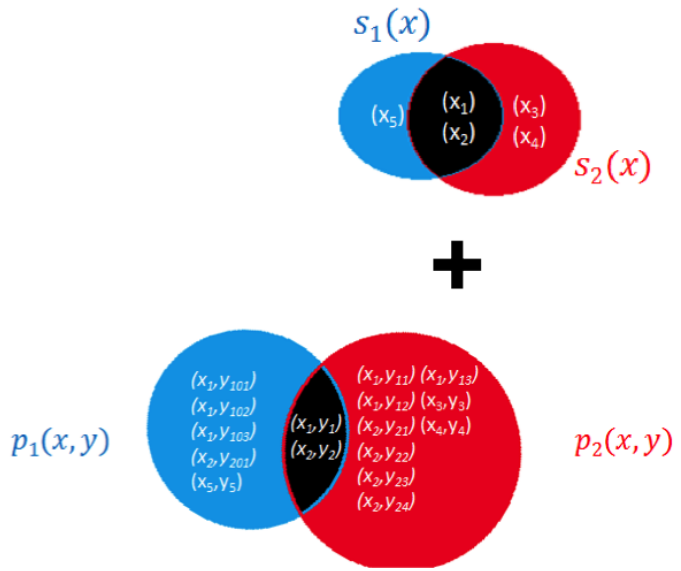
Determine Synonymity

- Measure the degree of synonymity
  - ***Triple overlap***
    overlap in the subject-object (argument) pairs
  - *Subject agreement*
    shared subjects of which sharedObject=true
  - *Cardinality ratio*
    do p_1 and p_2 have about the same objects per subject?
- Arbitrary cut-off threshold T_{minSyn}



Determine Synonymity

- Measure the degree of synonymity
  - *Triple overlap*
    overlap in the subject-object (argument) pairs
  - ***Subject agreement***
    shared subjects of which sharedObject=true
  - *Cardinality ratio*
    do p_1 and p_2 have about the same objects per subject?
- Arbitrary cut-off threshold T_{minSyn}

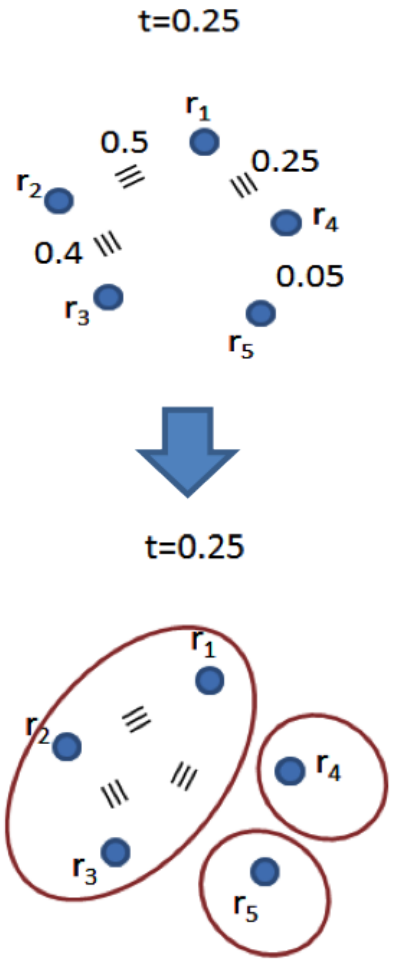


Determine Synonymity

- Measure the degree of synonymity
  - *Triple overlap*
    overlap in the subject-object (argument) pairs
  - *Subject agreement*
    shared subjects of which sharedObject=true
  - ***Cardinality ratio***
    do p_1 and p_2 have about the same objects per subject?
- Arbitrary cut-off threshold T_{minSyn}

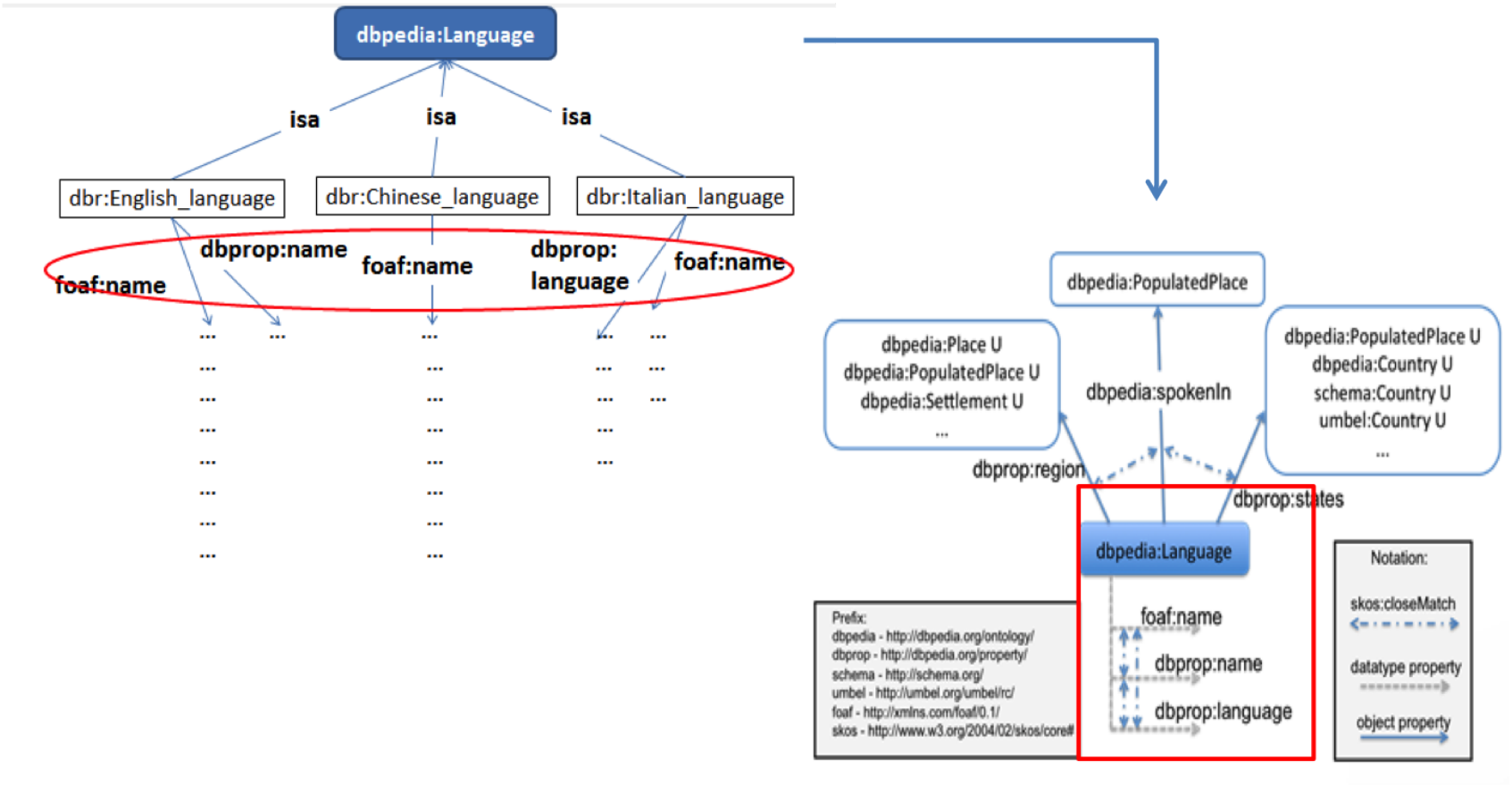
Clustering

- Rule-based agglomerative clustering
  - Minimum threshold T_{minSyn}
  - Minimum size of p (%triples covered for the concept)
  - Distance from cluster



For more details refer to [Zhang et al., 2013]

Selecting Properties for SKP



Selecting Properties for SKP

- Use frequency of property usage, treating clusters as one property
  - Calculate frequency, then use cut-off threshold
    - Absolute: cover at least # of triples
    - Fractional: cover at least % of triples for a class
    - Normalised: cover at least % of the average # of triples per property for the class
- *skos:closeMatch* added between synonymous properties

Evaluation

1. SKP observation
   - descriptive capability of SKPs
     - characterising underlying data
     - giving access to data

2. Query expansion
   - worthiness of“synonymous” properties
     - improve on retrieval coverage
     - without introducing erroneous data

SKP observation: evaluation settings

- Selecting properties for SKP based on frequency in data usage
  - Absolute: cover at least # of triples
  - Fractional: cover at least % of triples for a class
  - **Normalised**: cover at least % of the average # of triples per property for the class

For more details refer to [Blomqvist et al., 2013]

SKP observation: evaluation results

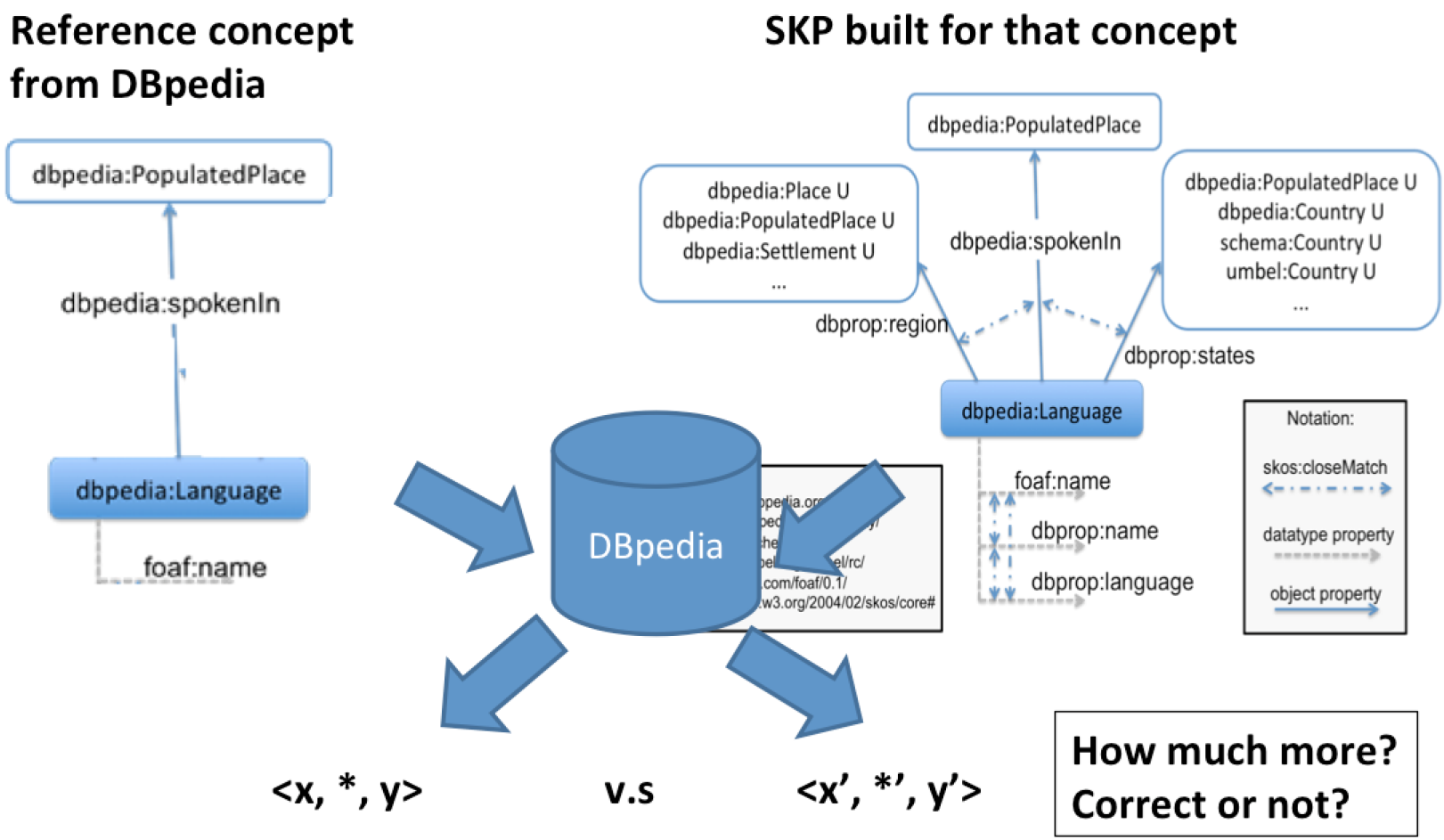
	Min	Average	Max
Number of included properties	31	107	436
Percentage of included properties	12%	27%	38%
Percentage of data triples covered	88%	94%	97%

SKP observation: evaluation results

	Min	Average	Max
Number of included properties	31	107	436
Percentage of included properties	12%	27%	38%
Percentage of data triples covered	88%	94%	97%

On average, we can **drop 73% of properties** of a class yet still able to **access 94% of triples**

Query expansion: evaluation settings

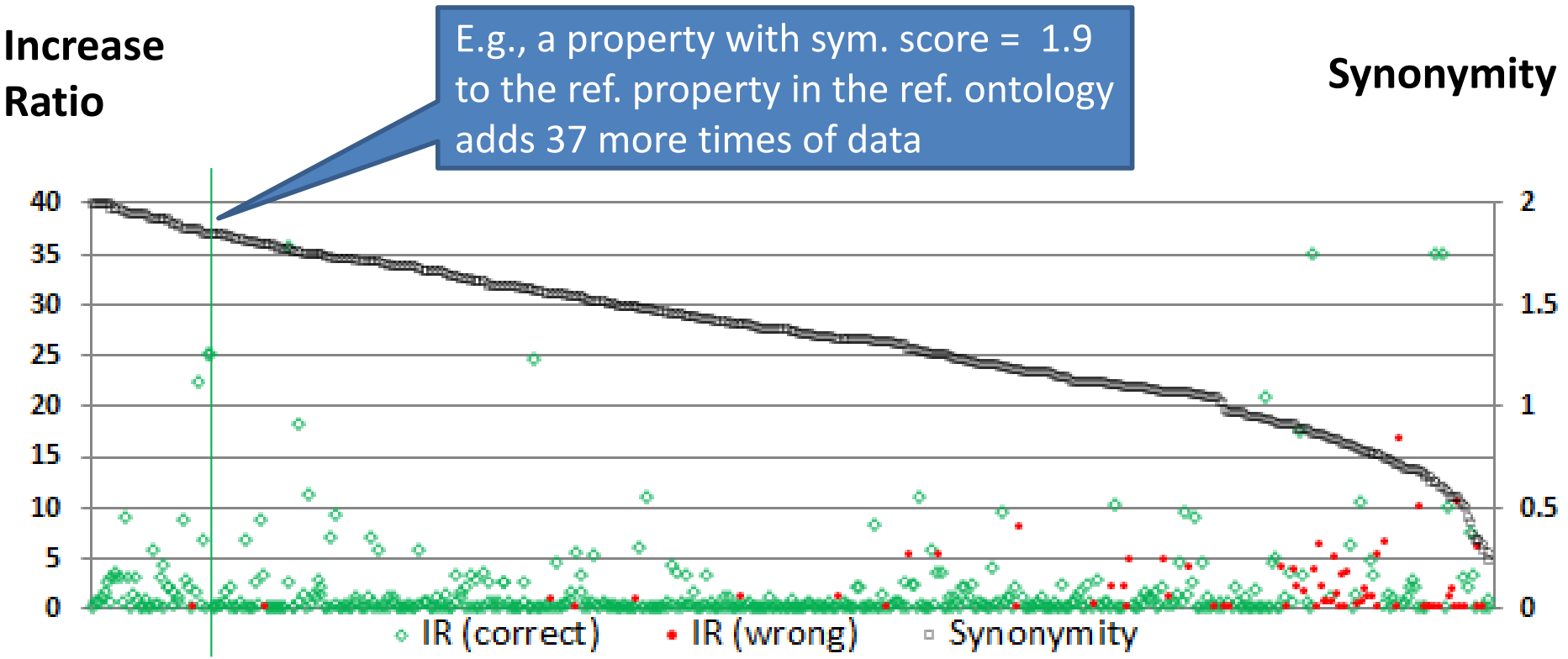


Query Expansion: Evaluation Settings

- Rule-based agglomerative clustering
  - Minimum threshold T_{minSyn} =0.1
  - Minimum size of p (%triples covered for the concept) = 0.01%
  - Distance from cluster = 0.6
- Correctness
  - 4 annotators look at each added property in SKP and decide if either it is a correct synonymous property to the original defined in the reference concept ontology
  - corresponding triples retrieved by the property marked accordingly

Query Expansion: Evaluation Results

- How much more triples are added due to SKP and are they correct?



SKP benefits and contributions

- context dependent
  - describe the usage and meaning of properties in the context of a particular class
- completely automatic
- general across vocabularies/datasets
- can be generated offline
  - can be used efficiently at run time
- their structure allows accurate query expansion

Further reading I

Blomqvist, E., Zhang, Z., Gentile, A. L., Augenstein, I., and Ciravegna, F. (2013).
Statistical knowledge patterns for characterizing linked data.
In *Proceedings of the 4th Workshop on Ontology and Semantic Web Patterns (WOP 2013)*, page to appear.

Nuzzolese, A. G., Gangemi, A., Presutti, V., and Ciancarini, P. (2011).
Encyclopedic knowledge patterns from wikipedia links.
In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, pages 520–536, Berlin, Heidelberg. Springer-Verlag.

Zhang, Z., Gentile, A. L., Augenstein, I., Blomqvist, E., and Ciravegna, F. (2013).
Mining equivalent relations from linked data.
In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 289–293, Sofia, Bulgaria. Association for Computational Linguistics.