



UNIVERSITY OF
Southampton



Preliminary Results in Tag Disambiguation using DBpedia

Andrés García-Silva[†], Martin Szomszor[‡], Harith Alani[‡], Oscar Corcho[†]

[†] {hgarcia, ocorcho}@fi.upm.es

Facultad de Informática

Universidad Politécnica de Madrid

Campus de Montegancedo s/n

28660 Boadilla del Monte, Madrid, Spain

[‡] {mns2, h.alani}@ecs.soton.ac.uk

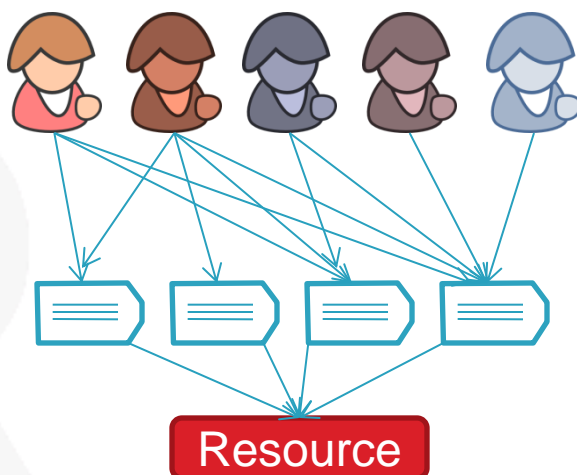
School of Electronics and

Computer Science,

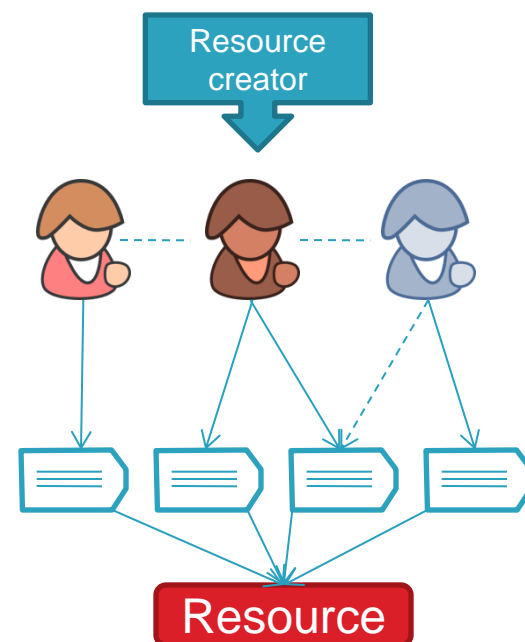
University of Southampton,

SO16 1BJ, UK.

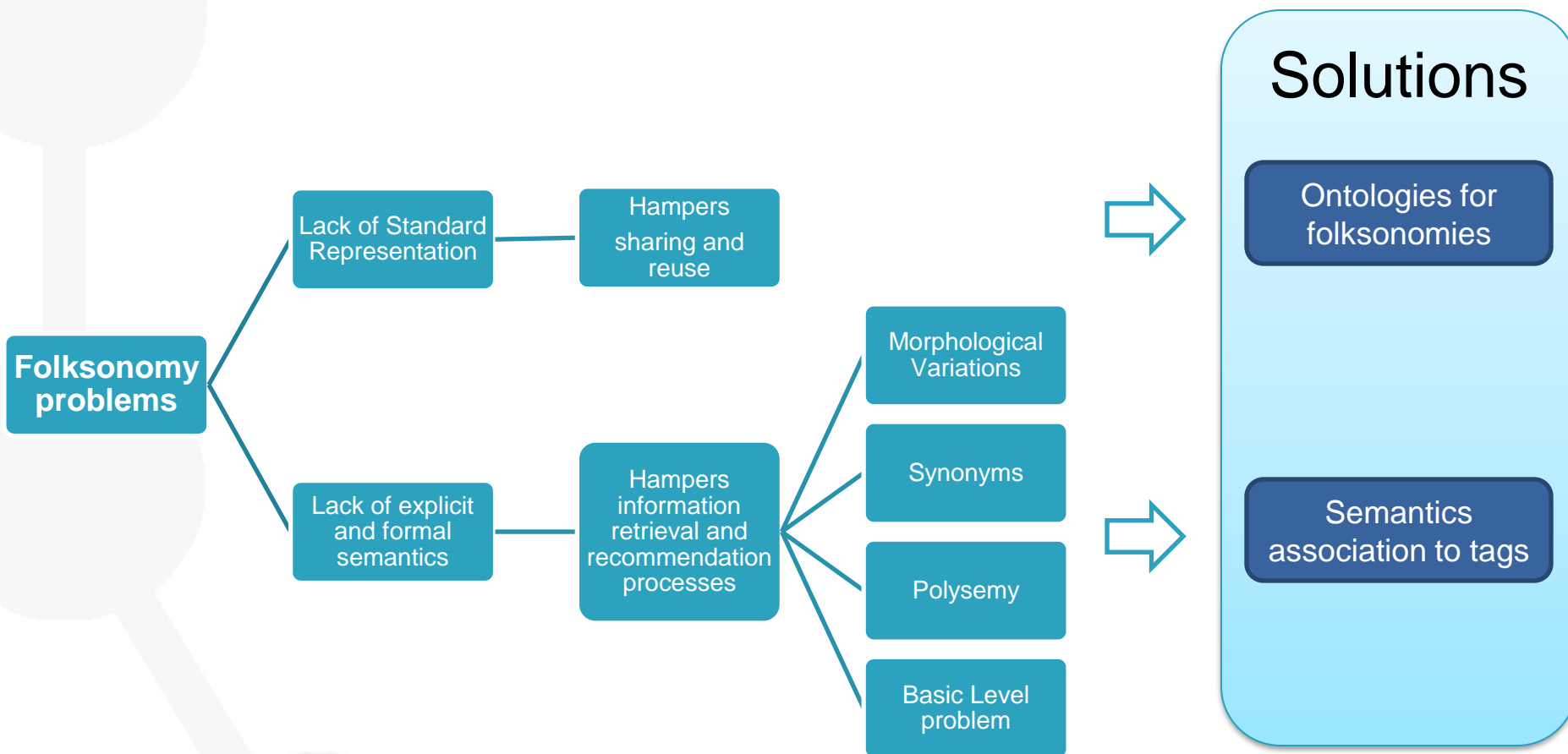
Broad folksonomy*

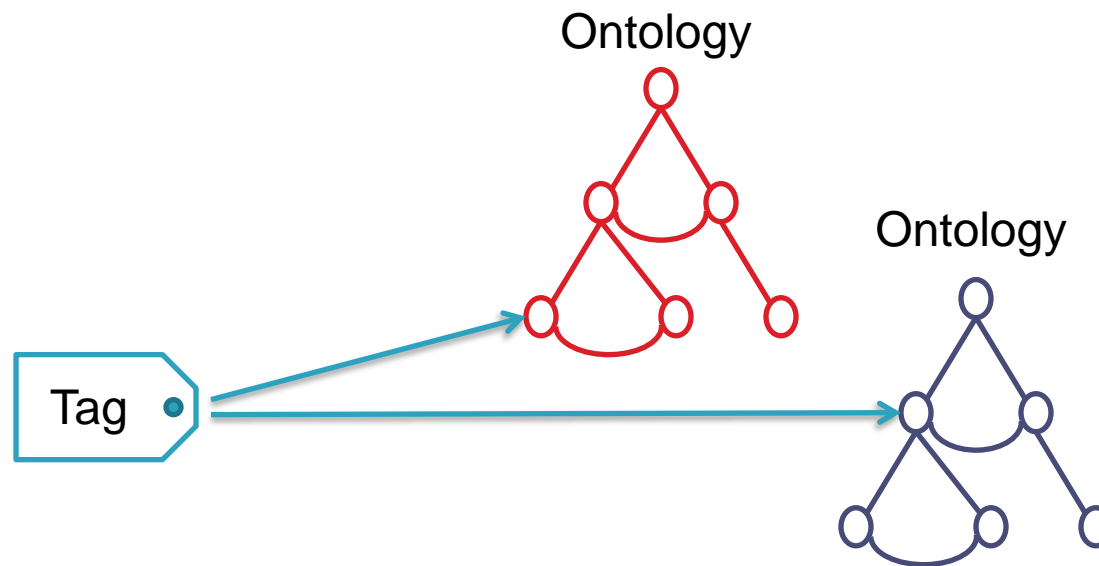


Narrow folksonomy*



* www.vanderwal.net





Semantics association to tags

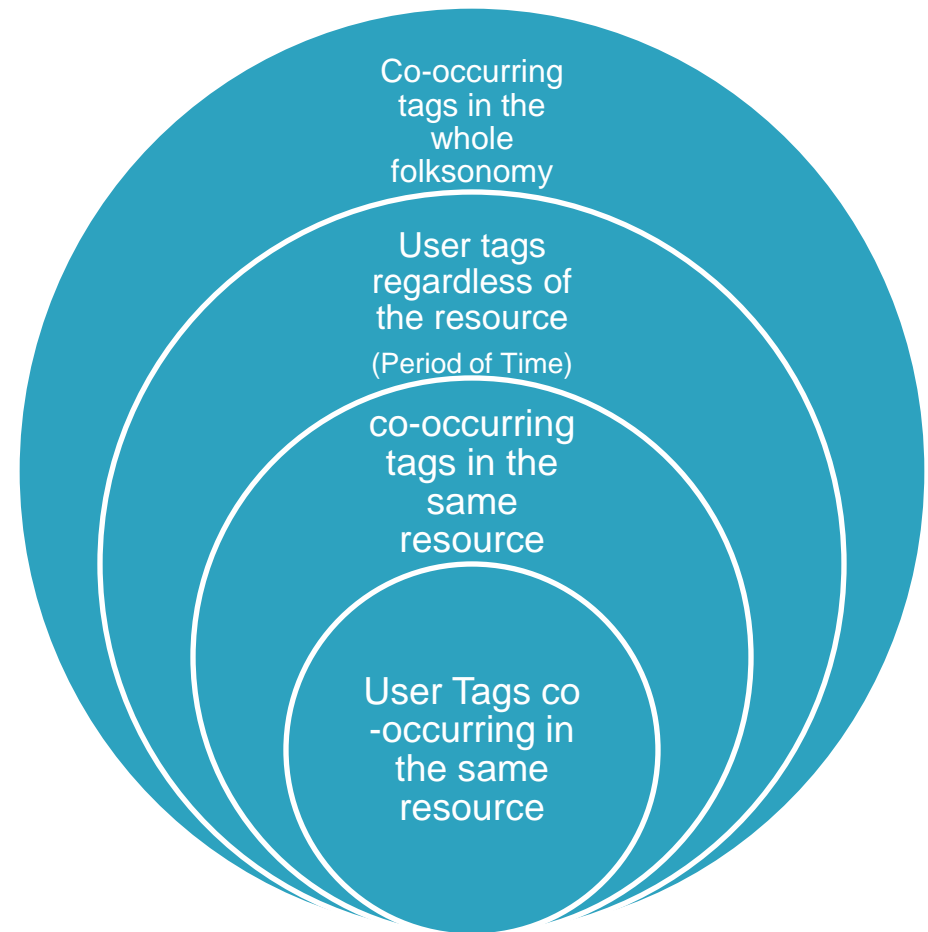


- First disambiguation approach relying on a dictionary (Lesk, 1998)
 - Definitions of the word to disambiguate & of each word in the context.
 - Context: The words appearing in the sentence
 - Definitions of the words in the context are compared against the definitions of the word to disambiguate.

Problems: When the definitions are short. (Sanderson, 2000)

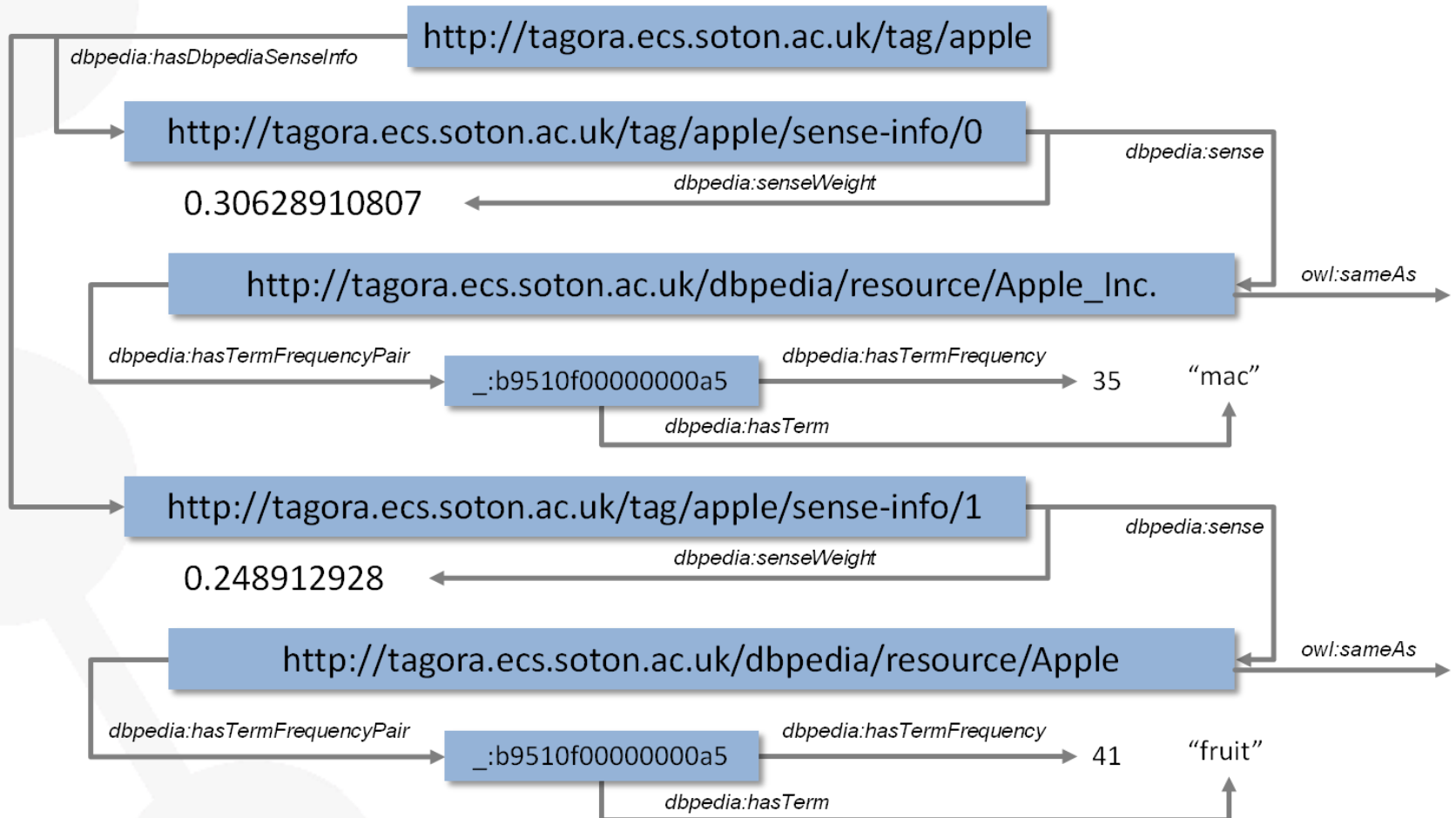
- Lesk, M., "They said true things, but called them by wrong names" – vocabulary problems in retrieval systems. in *Proc. 4th Annual Conference of the University of Waterloo Centre for the New OED* (1998)
- Sanderson, M., Retrieving with Good Sense. In *Information Retrieval* 2(1): 47-67 (2000)

- Contexts in folksonomies



- Linked data enabled service endpoint
 - Metadata about tags and their possible senses.
 - Wikipedia pages -> Disambiguation or Redirection links
 - Terms and frequencies
 - DBpedia entries related to each Wikipedia page.
 - Query using:
 - REST -> <http://tagora.ecs.soton.ac.uk/tag/apple>
 - SPARQL end-point.
 - Result: RDF document
- DBpedia coverage:
 - **2.6** million things, 213,000 people, 328,000 places, 57,000 music albums, 36,000 films, 20,000 companies.
 - Wordnet as of 2006 contains about 150,000 words
 - Named entity recognition
 - Classes, Instances, and semantic relations

Linked data representation of tag senses



- The algorithm selects among a set of candidate DBpedia entries, the one that describe better the meaning of an ambiguous tag according to its context.
- The candidate DBpedia entries and the tag context are represented as vectors.
- The tag context vector is compared against each DBpedia entry vector using the cosine of the angle as similarity measure.

$$Sim(V_{context}, V_{sense}) = \cos \theta = \frac{V_{context} \cdot V_{sense}}{|V_{context}| |V_{sense}|}$$

- The most similar DBpedia entry is selected as the one representing the meaning of the analyzed tag

- Tagging activity:
 - User u has tagged the resource $r = \text{http://www.nature.com}$ with the tags $nature$, $news$, $science$.
- **Context**(u , $nature$, r) = { $nature$, $news$, $science$ }
- **Senses**($Nature$) = { $dbpedia:Nature$, $dbpedia:Nature_journal$ }
- $\text{Terms}(dbpedia:Nature) = \{(life, 62), (nature, 46), (earth, 32)\}$
- $\text{Terms}(dbpedia:Nature_journal) = \{(nature, 77), (science, 29), (scientific, 25)\}$
- **Voc**($nature$) = { $life$, $nature$, $earth$, $science$, $scientific$ }
- $V_{\text{context}} = (0, 1, 0, 1, 0)$
- $V_{\text{nature}} = (62, 46, 32, 0, 0)$
- $V_{\text{nature(journal)}} = (0, 77, 0, 29, 25)$
- $\text{Sim}(V_{\text{context}}, V_{\text{nature}}) = 0,389$ $\text{Sim}(V_{\text{context}}, V_{\text{nature(journal)}}) = \mathbf{0,872}$

Some user x has tagged a picture r with the tags *ice*, *iceskating*, *nottingham*, and *skating*.



<i>ice</i>	
dbpedia/resource/Ice	0,911
dbpedia/resource/Ice_(comics)	0,735
<i>skating</i>	
dbpedia/resource/Artistic_roller_skating	0,671
dbpedia/resource/Figure_skating	0,569
dbpedia/resource/Freestyle_slalom_skating	0,000
dbpedia/resource/Ice_skating	0,893
dbpedia/resource/Road_skating	0,451
dbpedia/resource/Roller_skating	0,394
dbpedia/resource/Skateboarding	0,197
dbpedia/resource/Snowboarding	0,000
dbpedia/resource/Speed_skating	0,549
dbpedia/resource/Tour_skating	0,831
<i>nottingham</i>	
dbpedia/resource/East_Nottingham_Township,_Pennsylvania	0,000
dbpedia/resource/Elizabeth_I_of_England	0,000
dbpedia/resource/Nottingham	0,750
dbpedia/resource/Nottingham,_New_Hampshire	0,386
dbpedia/resource/Nottingham_Cooperative	0,524
dbpedia/resource/Nottingham_Township,_Harrison_County,_Ohio	0,000
dbpedia/resource/Nottingham_Township,_Pennsylvania	0,000
dbpedia/resource/Nottinghamshire	0,428
dbpedia/resource/Sheriff_of_Nottingham	0,640
dbpedia/resource/West_Nottingham_Township,_Pennsylvania	0,000

- Conclusions:
 - Inspired by IR techniques we have presented a tag disambiguation algorithm relying on DBpedia & Wikipedia information.
 - We have present different definitions of contexts for tagging activities as a way to ameliorate tagging data scarceness.
- Future Work
 - Test the approach in a large tag set with the different contexts.
 - Sophisticated similarity measures
 - Evaluation of the approach and contexts using *Precision* and *Recall*
 - Extract more context information from the tagged resource (text documents)
 - Test bed and Standard evaluation metrics
 - **Use DBpedia semantic information to evolve domain ontologies**
 - **Use DBpedia semantic information to improve searching and recommendation processes.**