

Conceptual Similarity: Why, Where, How

Michalis Sfakakis

Laboratory on Digital Libraries & Electronic Publishing,
Department of Archives and Library Sciences,
Ionian University, Greece

First Workshop on Digital Information Management, March 30-31, 2011
Corfu, Greece

Do we need similarity?

- Are the following objects similar?
 - (Similarity, SIMILARITY)
 - As character sequences, NO!
 - How do they differ?
 - As character sequences, but case insensitive, Yes!
 - As English words, Yes!
 - Same word! They have the same definition, written differently
-

Contents

- Introduction

- Disciplines

- How we measure similarity

 - Focus on Ontology Learning evaluation

Exploring similarity... more cases

- What about the similarity of the objects?
 - (1, a)
 - The first object is the number one and the second is the first letter of the English alphabet. Therefore, as the first is a number and the second is a letter, they are different!

 - But, conceptually... When both represent an order, e.g. a chapter, or a paragraph number, they are both representing the first object of the list, the first chapter, paragraph, etc. Therefore, they could be considered as being similar!
-

Results for an Information Need

Screenshot of Google search results for the query "databases". The browser address bar shows "http://www.google.com/search?". The left sidebar contains navigation options like "Everything", "Images", "Videos", "News", "Shopping", "Books", and "More". The main content area displays several search results:

- Database - Wikipedia, the free encyclopedia**: A database is a system intended to organize, store, and retrieve large amounts of data easily. It consists of an organized collection of data for one or ... Management system - Design - Relational database - Comparison of database tools en.wikipedia.org/wiki/Database - Cached - Similar
- EBSCOhost - world's foremost premium research database service**: EBSCOhost (ebSCOhost.com) serves thousands of libraries and other institutions with premium content in every subject area. Free LISTA: LibraryResearch.com. search.ebSCOhost.com/ - Cached - Similar
- About Databases: Microsoft Access, SQL Server, Oracle and More!**: 24 Mar 2011 ... Your About.com Guide to Databases provides a comprehensive look at database systems with feature articles, the Net's best collection of ... databases.about.com/ - Cached - Similar
- What is database? - A Word Definition From the Webopedia Computer ...**: This page describes the term database and lists other pages on the Web where you can find additional information. www.webopedia.com/TERM/D/database.html - Cached
- Databases - United Nations**: This is the Databases page of the United Nations website. Here you will find links to various information resources. www.un.org/en/databases/index.shtml - Cached
- Bioinformatics Databases | EBI**: EB-eye, ebsearch, search+the+ebi, UniProt, UniRef, UniParc, Universal Protein Resource, EBI, EMBL, bioinformatics, molecular, genetics, software, databases ... www.ebi.ac.uk/Databases/ - Cached - Similar
- BBC - GCSE Bitesize - Data, information and databases**: A GCSE revision and recap resource for ICT Data, information and databases. www.bbc.co.uk > Home > ICT - Cached - Similar - Add to iGoogle
- MySQL :: The world's most popular open source database**: The software's official homepage with news, downloads and documentation. www.mysql.com/ - Cached - Similar
- Database**: 13 Jan 2009 ... Entrez is the text-based search and retrieval system used at NCBI for all of the major databases, including PubMed, Nucleotide and Protein ... www.ncbi.nlm.nih.gov/Database/ - Cached
- AustLII: AustLII Databases**: You are here: AustLII >> AustLII Databases. Cth | ACT | NSW | NT | Qld | SA | Tas | Vic | WA | Norfolk Island | New Zealand Journals | Law Reform | Special ... www.austlii.edu.au/databases.html - Cached - Similar

Screenshot of Bing search results for the query "databases". The browser address bar shows "http://www.bing.com/". The left sidebar contains navigation options like "RELATED SEARCHES", "Free Database", "Business Databases", "Types of Databases", "Free Online Database", "Definition of Database", "Microsoft Access Database", "Database Normalization", and "Database Icon". The main content area displays several search results:

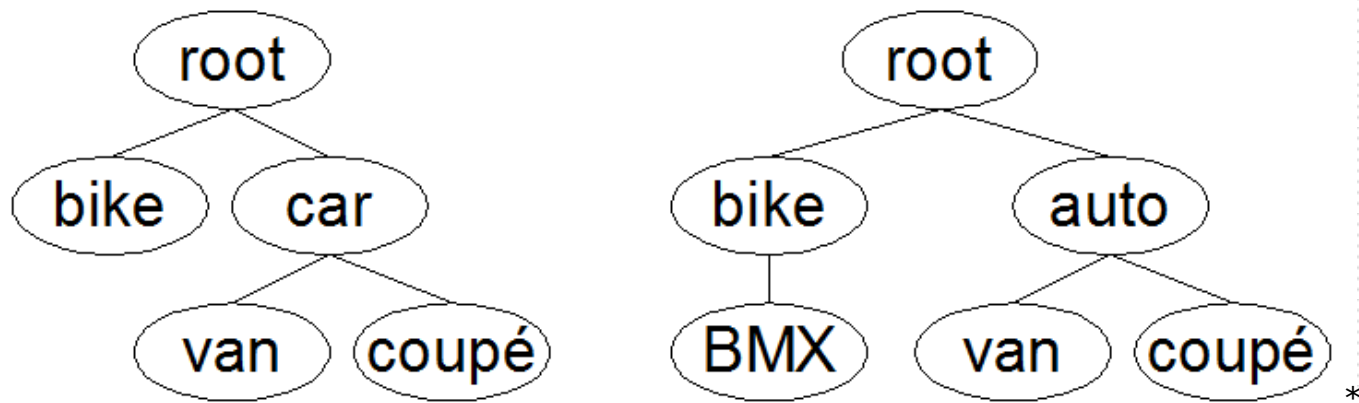
- Database - Wikipedia, the free encyclopedia**: Architecture · Database management ... · Types · Models Galindo, J.; Urrutia, A.; Piattini, M. Fuzzy Databases: Modeling, Design and Implementation (FSQL guide). Idea Group Publishing Hershey, USA, 2006. en.wikipedia.org/wiki/Database
- About Databases: Microsoft Access, SQL Server, Oracle and More!**: Your About.com Guide to Databases provides a comprehensive look at database systems with feature articles, the Net's best collection of links, forums and newsletters! databases.about.com
- database: Definition from Answers.com**: Galindo, J.; Urrutia, A.; Piattini, M. Fuzzy Databases: Modeling, Design and Implementation (FSQL guide). Idea Group Publishing Hershey, USA, 2006. www.answers.com/topic/database
- EBSCOhost - world's foremost premium research database ...**: Important User Information: Remote access to EBSCO's databases is permitted to patrons of subscribing institutions accessing from remote locations for personal, non ... search.ebSCOhost.com
- NoSQL - Wikipedia, the free encyclopedia**: History · Architecture · Taxonomy · Others Academics and papers typically refer to these databases as structured storage a term that would include classic relational databases as a subset. en.wikipedia.org/wiki/NoSQL
- Databases - United Nations**: This is the Databases page of the United Nations website. Here you will find links to various information resources. www.un.org/en/databases/index.shtml
- Microsoft Access Database - Software, training products, ebooks ...**: Our Microsoft Access Database site gives you all the tools and training you need to take your business forward. access-databases.com
- Los Angeles Public Library | Databases**: Los Angeles Public Library's Online Databases ... Find Articles & Information. Scroll down to select an individual database from the list below. databases.lapl.org
- Database**: Entrez is the text-based search and retrieval system used at NCBI for all of the major databases, including PubMed, Nucleotide and Protein Sequences, Protein ...

□ How similar are the Results? Which one to select?

Comparing Concepts

- ... again, how similar are the following objects?
 - (Disease, Illness)
 - As English words, or as character sequences they are not similar!
 - How do they differ?
 - As synonymous terms in a Thesaurus, they are both representing the same concept. (related with the *equivalency* relationship)
-

Comparing Hierarchies



□ How similar...

- ... is the node *car* from the left hierarchy to the node *auto* from the right hierarchy?
- ... are the nodes *van* from both hierarchies?
- ... is the above hierarchies?

* [Dellschaft and Staab, 2006]

... so, what similarity is?

- Similarity is a context dependent concept
- Merriam-Webster's Learner's dictionary defines similarity as*:
 - A quality that makes one person or thing like another
 - ... and similar, having characteristics in common
- Therefore, the context and the characteristics in common are required in order to specify and measure similarity

* <http://www.learnersdictionary.com/search/similarity>

Where the concept of similarity is encountered

- ... Similarity is a context dependent concept

 - Machine learning
 - Ontology Learning
 - Schema & Ontology Matching and Mapping
 - Clustering
 - IR
 - ... in any evaluation concerning the results of a pattern recognition algorithm

 - Vital part of the Semantic Web development
-

Precision & Recall in IR, measuring similarity between answers

- Let C be the result set for a query (the retrieved documents, i.e. the *Computed* set)

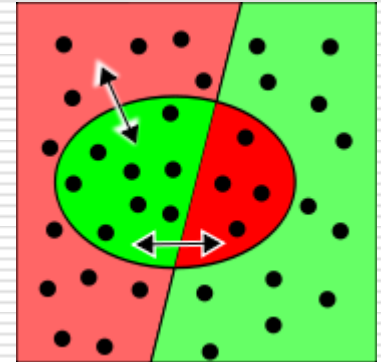
- Also, we need to know the correct results for the query (all the relevant documents, the *Reference* set)
 - *Precision*: is the fraction of retrieved documents that are relevant to the search
 - *Recall*: is the fraction of the documents that are relevant to the query that are successfully retrieved

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad \text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

... Precision & Recall, a way to measure similarity

- *Precision & Recall* are two widely used metrics for evaluating the correctness of a pattern recognition algorithm

- *Recall* and *Precision* depend on the outcome (oval) of a pattern recognition algorithm and its relation to all relevant patterns (left) and the non-relevant patterns (right).
The more correct results (green), the better.
 - *Precision*: horizontal arrow.
 - *Recall*: diagonal arrow.



Precision & Recall, once more

□ Precision

■ $P = |R \cap C| / |R|$

□ Recall

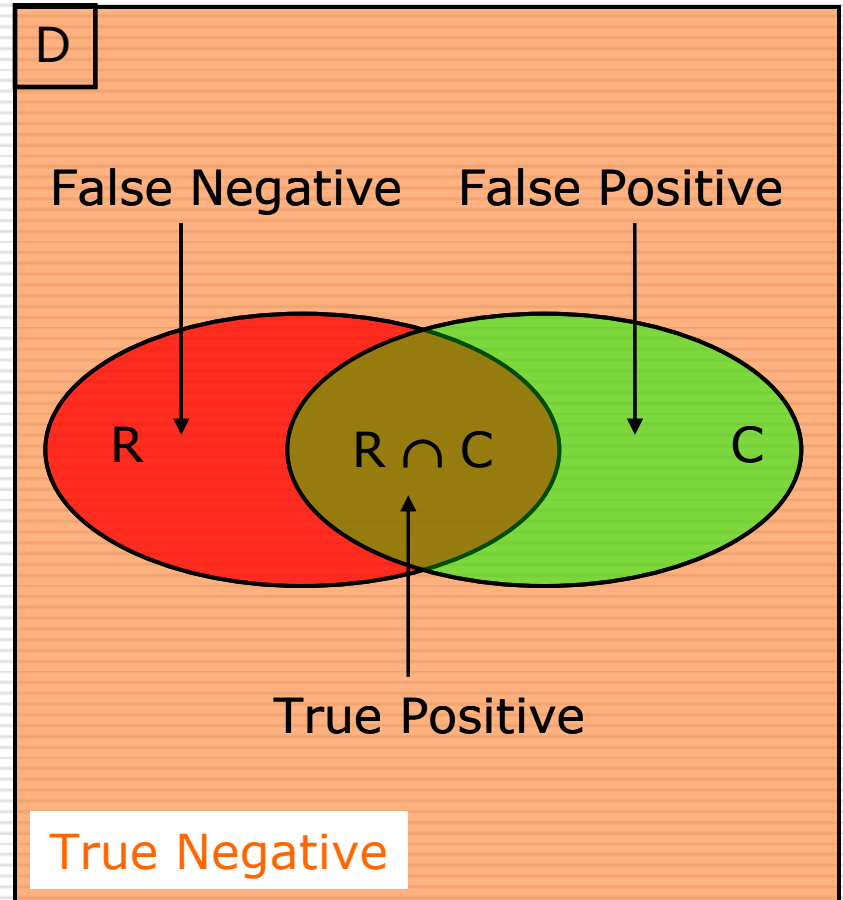
■ $R = |R \cap C| / |C|$

□ $TP = R \cap C$

□ $TN = D - (R \cup C)$

□ $FN = R - C$

□ $FP = C - R$



Overall evaluation, combining *Precision* & *Recall*

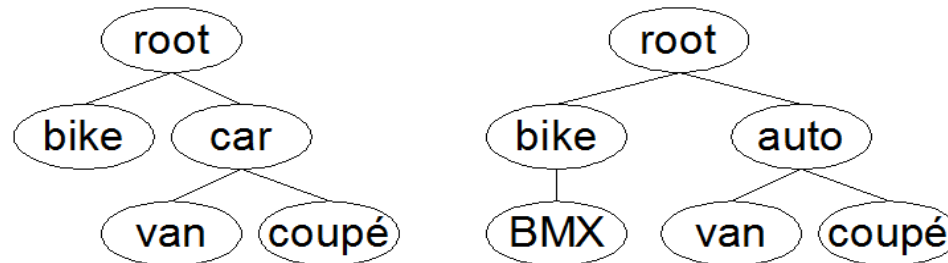
- Given *Precision* & *Recall*, *F-measure* could combines them for an overall evaluation

 - Balanced *F-measure* (*P* & *R* are evenly weighted)
 - $F_1 = 2*(P*R)/(P+R)$

 - Weighted *F-measure*
 - $F_b = (1+b^2)*(P*R)/(b^2*P+R)$, *b* non-zero

 - F_1 (*b*=2) weights recall twice as much as precision
 - $F_{0.5}$ (*b*=0.5) weights precision twice as much as recall
-

Measuring Similarity, Comparing two Ontologies

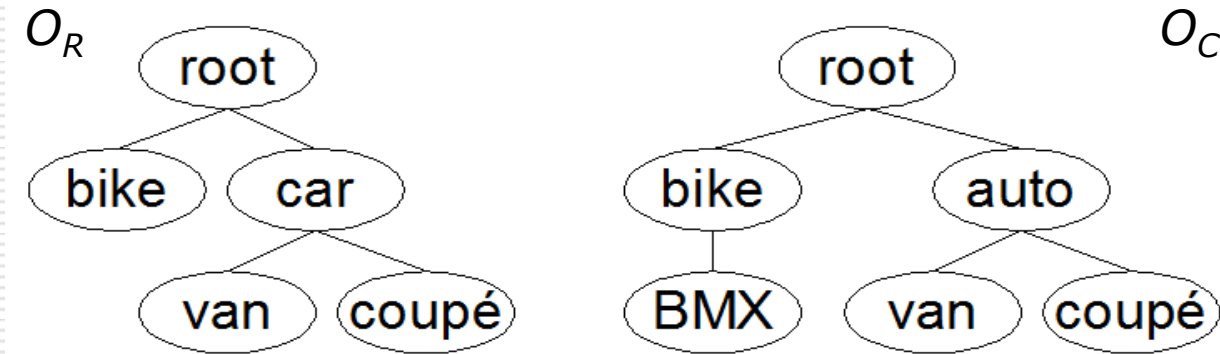


- A simplified definition of a core ontology*:
 - The structure $O := (C, root, \leq_C)$ is called a core ontology. C is a set of concept identifiers and $root$ is a designated root concept for the partial order \leq_C on C . This partial order is called concept hierarchy or taxonomy. The equation $\forall c \in C : c \leq_C root$ holds for this concept hierarchy.

- Levels of comparison
 - Lexical, how terms are used to convey meanings
 - Conceptual, which conceptual relations exist between terms
 - ...

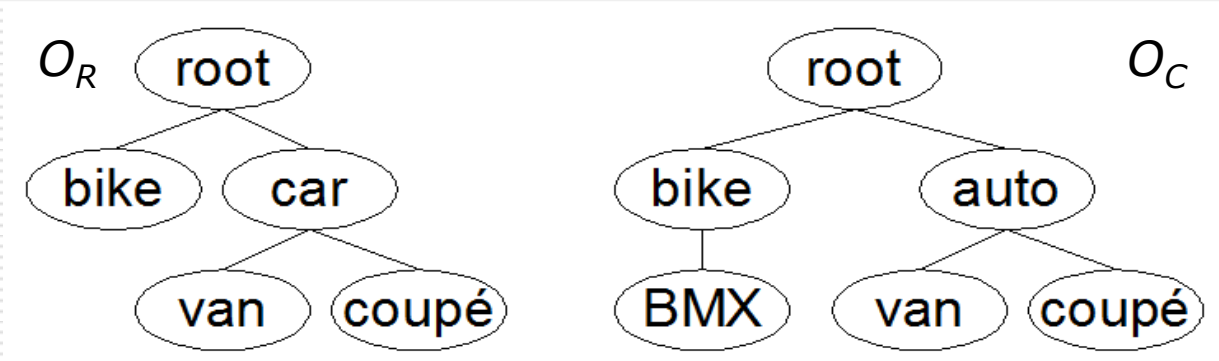
* [Dellschaft and Staab, 2006]

Gold Standard based Evaluation of Ontology Learning



- Given a pre-defined ontology
 - The so-called *Gold Standard* or *Reference*
 - Compare the *Learned (Computed) Ontology* with the *Gold Standard*
-

Measuring Similarity - Lexical Comparison Level - LP, LR



□ *Lexical Precision & Lexical Recall*

■ $LP(O_C, O_R) = |C_C \cap C_R| / |C_C|$

■ $LR(O_C, O_R) = |C_C \cap C_R| / |C_R|$

□ The lexical precision and recall reflect how good the learned lexical terms C_C cover the target domain C_R

□ For the above example $LP=4/6=0.67$, $LR=4/5=0.8$

Measuring Similarity, Lexical Comparison Level - aSM

- *Average String Matching*, using edit distance
 - *Levenshtein distance*, the most common definition for edit distance, measures the minimum number of token insertions, deletions and substitutions required to transform one string into an other

 - For example*, the *Levenshtein distance* between "*kitten*" and "*sitting*" is 3 (there is no way to do it with fewer than three edits)
 - **k**itten → **s**itten (substitution of 's' for 'k')
 - sit**t**en → sitt**i**n (substitution of 'i' for 'e')
 - sittin → sittin**g** (insertion of 'g' at the end).

* Wikipedia: http://en.wikipedia.org/wiki/Levenshtein_distance

Measuring Similarity, Lexical Comparison Level – String Matching

□ *String Matching* measure (SM), given two lexical entries L_1, L_2

$$SM(L_i, L_j) := \max \left(0, \frac{\min(|L_i|, |L_j|) - \text{ed}(L_i, L_j)}{\min(|L_i|, |L_j|)} \right) \in [0, 1]$$

- Weights the number of the required changes against the shorter string
- 1 stands for perfect match, 0 for bad match

□ *Average SM*

- *Asymmetric, determines the extent to which \mathcal{L}_1 (target) is covered by \mathcal{L}_2 (source)*

$$\overline{SM}(\mathcal{L}_1, \mathcal{L}_2) := \frac{1}{|\mathcal{L}_1|} \sum_{L_i \in \mathcal{L}_1} \max_{L_j \in \mathcal{L}_2} SM(L_i, L_j)$$

Measuring Similarity, Lexical Comparison Level - *RelHit*

□ Relative Number of Hits

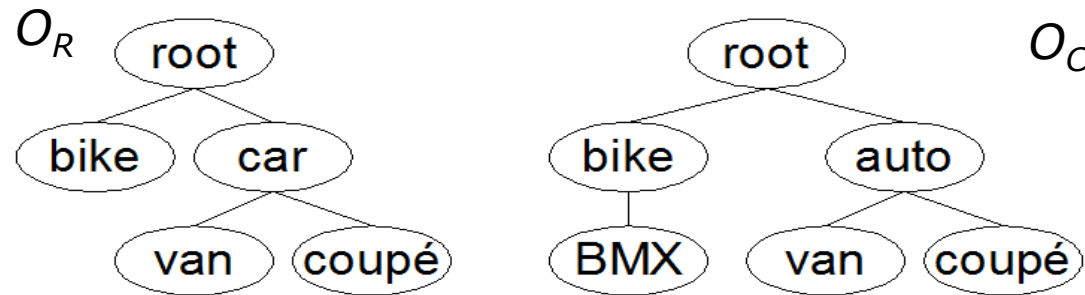
$$\text{RelHit}(\mathcal{L}_1, \mathcal{L}_2) := \frac{|\mathcal{L}_1 \cap \mathcal{L}_2|}{|\mathcal{L}_1|}$$

- *RelHit* actually express Lexical Precision
 - *RelHit* Compared to average String Matching
 - *Average SM* reduces the influences of string pseudo-differences (e.g. singular vs. plurals)
 - *Average SM* may introduce some kind of noise, e.g. "power", "tower"
-

Measuring Similarity, Conceptual Comparison Level

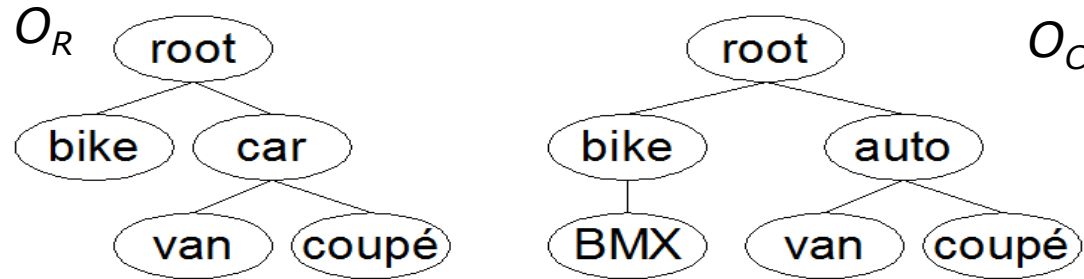
- ❑ Conceptual level compares semantic structure of ontologies
 - ❑ Conceptual structures are constituted by Hierarchies, or by Relations
 - ❑ How to compare two hierarchies?
 - ❑ How do the positions of concepts influence similarity of Hierarchies?
 - ❑ What measures to use?
-

Measuring Similarity, Conceptual Comparison Level



- Local measures compare the positions of two concepts based on characteristics extracts from the concept hierarchies they belong to
 - Some characteristic extracts
 - Semantic Cotopy (sc)
 - $sc(c, O) = \{c_i | c_i \in C \wedge (c_i \preceq c \vee c \preceq c_i)\}$
 - Common Semantic Cotopy (csc)
 - $csc(c, O_1, O_2) = \{c_i | c_i \in C_1 \cap C_2 \wedge (c_i <_1 c \vee c <_1 c_i)\}$
-

Measuring Similarity, Conceptual Comparison Level – sc



□ Semantic Cotopy

- $sc(c, O) = \{c_i | c_i \in C \wedge (c_i \leq c \vee c \leq c_i)\}$

□ Semantic Cotopy examples

- $sc(\text{"root"}, O_R) = \{\text{root}, \text{bike}, \text{car}, \text{van}, \text{coupé}\}$

- $sc(\text{"root"}, O_C) = \{\text{root}, \text{bike}, \text{auto}, \text{BMX}, \text{van}, \text{coupé}\}$

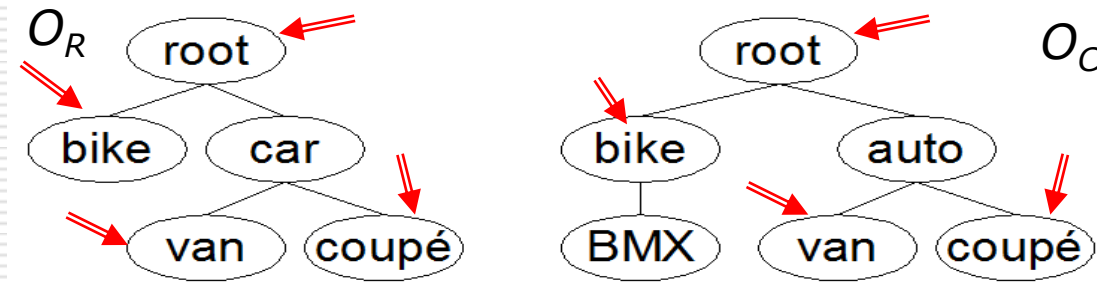
- $sc(\text{"bike"}, O_R) = \{\text{root}, \text{bike}\}$

- $sc(\text{"bike"}, O_C) = \{\text{root}, \text{bike}, \text{BMX}\}$

- $sc(\text{"car"}, O_R) = \{\text{root}, \text{car}, \text{van}, \text{coupé}\}$

- $sc(\text{"auto"}, O_C) = \{\text{root}, \text{auto}, \text{van}, \text{coupé}\}$

Measuring Similarity, Conceptual Comparison Level – csc



□ Common Semantic Cotopy

- $csc(c, O_1, O_2) = \{c_i | c_i \in C_1 \cap C_2 \wedge (c_i <_1 c \vee c <_1 c_i)\}$

□ Common Semantic Cotopy examples

- $C_1 \cap C_2 = \{\text{root, bike, van, coupé}\}$

- $csc(\text{"root"}, O_R, O_C) = \{\text{bike, van, coupé}\}$

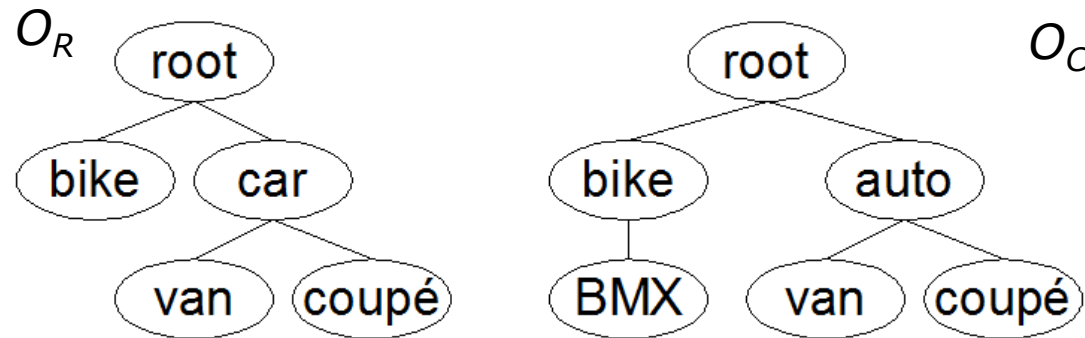
- $csc(\text{"root"}, O_C, O_R) = \{\text{bike, van, coupé}\}$

- $csc(\text{"bike"}, O_R, O_C) = \{\text{root}\}, csc(\text{"bike"}, O_C, O_R) = \{\text{root}\}$

- $csc(\text{"car"}, O_R, O_C) = \{\text{root, van, coupé}\}, csc(\text{"car"}, O_C, O_R) = \emptyset$

- $csc(\text{"auto"}, O_C, O_R) = \{\text{root, van, coupé}\}, csc(\text{"auto"}, O_R, O_C) = \emptyset$

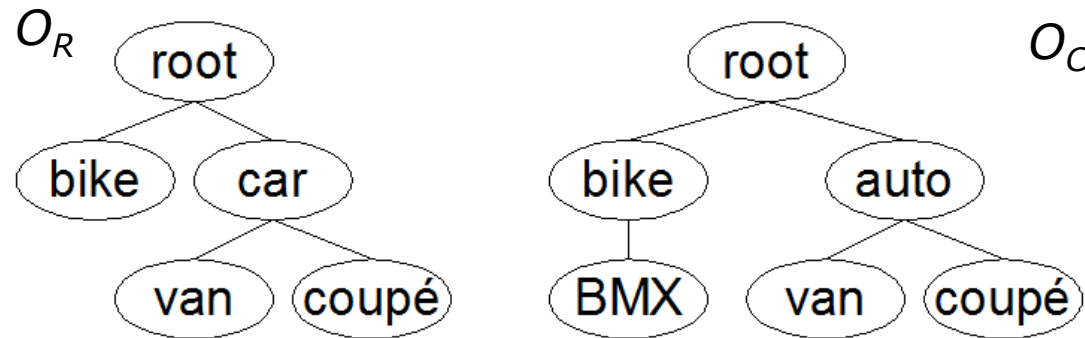
Measuring Similarity, Conceptual Comparison Level – local measures tp , tr



- Local *taxonomic precision* using characteristic extracts
 - $tp_{ce}(c_1, c_2, O_C, O_R) = |ce(c_1, O_C) \cap ce(c_1, O_R)| / |ce(c_1, O_C)|$

 - Local *taxonomic recall* using characteristic extracts
 - $tr_{ce}(c_1, c_2, O_C, O_R) = |ce(c_1, O_C) \cap ce(c_1, O_R)| / |ce(c_1, O_R)|$
-

Measuring Similarity, Conceptual Comparison Level – local measures tp

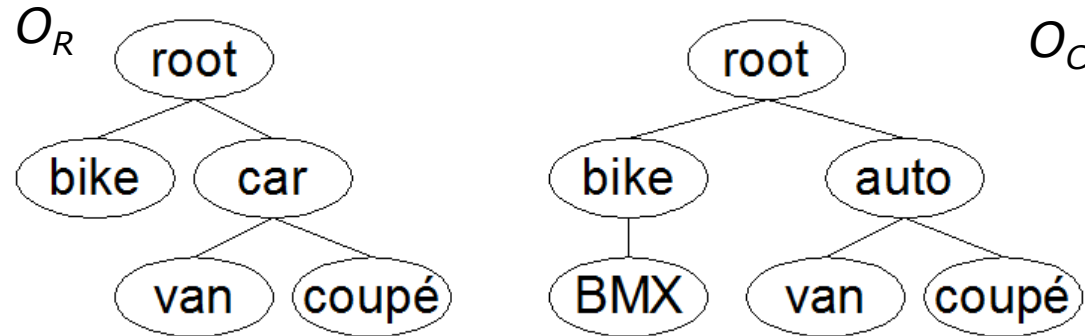


□ *Local taxonomic precision* examples using *sc*

- $sc(\text{"bike"}, O_R) = \{\text{root}, \text{bike}\},$
 $sc(\text{"bike"}, O_C) = \{\text{root}, \text{bike}, \text{BMX}\}$

- $tp_{sc}(\text{"bike"}, \text{"bike"}, O_C, O_R) = |\{\text{root}, \text{bike}\}| / |\{\text{root}, \text{bike}, \text{BMX}\}|,$
 $tp_{sc}(\text{"bike"}, \text{"bike"}, O_C, O_R) = 2/3 = 0.67$

Measuring Similarity, Conceptual Comparison Level – local measures tp

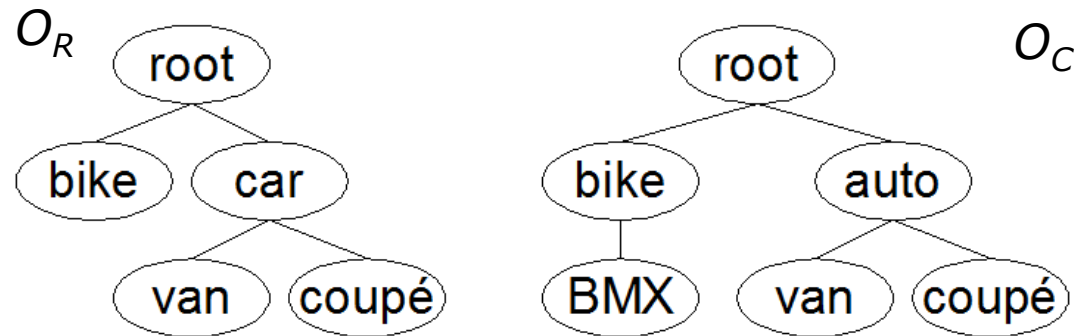


□ *Local taxonomic precision* examples using *sc*

- $sc(\text{"car"}, O_R) = \{\text{root}, \text{car}, \text{van}, \text{coupé}\}$,
 $sc(\text{"auto"}, O_C) = \{\text{root}, \text{auto}, \text{van}, \text{coupé}\}$

- $tp_{sc}(\text{"car"}, \text{"auto"}, O_C, O_R) =$
 $|\{\text{root}, \text{van}, \text{coupé}\}| / |\{\text{root}, \text{auto}, \text{van}, \text{coupé}\}|,$
 $tp_{sc}(\text{"car"}, \text{"auto"}, O_C, O_R) = 3/4 = 0.75$
-

Measuring Similarity, Conceptual Comparison Level – comparing Hierarchies



□ Global Taxonomic Precision (TP)

$$TP(\mathcal{O}_C, \mathcal{O}_R) := \frac{1}{|\mathcal{C}_C|} \sum_{c \in \mathcal{C}_C} \begin{cases} tp(c, c, \mathcal{O}_C, \mathcal{O}_R) & \text{if } c \in \mathcal{C}_R \\ \max_{c' \notin \mathcal{C}_R} tp(c, c', \mathcal{O}_C, \mathcal{O}_R) & \text{if } c \notin \mathcal{C}_R \end{cases}$$

Labels in the diagram:

- local taxonomic precision (pointing to the tp function)
- concept set (pointing to \mathcal{C}_C)
- estimation (pointing to the summation and the piecewise function)

Measuring Similarity, Conceptual Comparison Level – Overall evaluation

- ... again *F-measure*, but now using *Global Taxonomic Precision (TP)* and *Global Taxonomic Recall (TR)*
 - Balanced Taxonomic *F-measure* (*TP* & *TR* are evenly weighted)
 - $TF_1 = 2*(TP*TR)/(TP+TR)$
 - Weighted *TF-measure*
 - $TF_b = (1+b^2)*(TP*TR)/(b^2*TP+TR)$, *b* non-zero
 - TF_1 (*b*=2) weights recall twice as much as precision
 - $TF_{0.5}$ (*b*=0.5) weights precision twice as much as recall
-

Measuring Similarity, Conceptual Comparison Level – Taxonomic Overlap

- *Global Taxonomic Overlap... based on local taxonomic overlap (TO)*

$$\overline{TO}(O_1, O_2) = \frac{1}{|C_1|} \sum_{c \in C_1} TO(c, O_1, O_2)$$

$$TO(c, O_1, O_2) = \begin{cases} TO'(c, O_1, O_2) & \text{if } c \in C_2 \\ TO''(c, O_1, O_2) & \text{if } c \notin C_2 \end{cases}$$

$$TO'(c, O_1, O_2) := \frac{|SC(c, O_1, O_2) \cap SC(c, O_2, O_1)|}{|SC(c, O_1, O_2) \cup SC(c, O_2, O_1)|}$$

$$TO''(c, O_1, O_2) := \max_{c' \notin C_2} \frac{|SC(c, O_1, O_2) \cap SC(c', O_2, O_1)|}{|SC(c, O_1, O_2) \cup SC(c', O_2, O_1)|}$$

References & Further Reading

- Dellschaft, Klaas and Staab, Steffen (2006) : On How to Perform a Gold Standard Based Evaluation of Ontology Learning. In: I. Cruz et al. (Eds.) ISWC 2006. LNCS 4273, pp. 228–241. Springer, Heidelberg
 - Maedche, A., Staab, S. (2002): Measuring similarity between ontologies. In: Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW-2002). Siguenza, Spain
 - Staab, S. and Hotho, A.: Semantic Web and Machine Learning Tutorial (available at <http://www.uni-koblenz.de/~staab/Research/Events/ICML05tutorial/icml05tutorial.pdf>)
 - Bellahsene, Z., Bonifati, A., Rahm, E. (Eds.) (2011): Schema Matching and Mapping, Heidelberg, Springer- Verlag, ISBN 978-3-642-16517-7.
-

End of tutorial!

□ Thanks for your attention!

□ Michalis Sfakakis
