UMLS-Interface and UMLS-Similarity:

Open Source Software for Measuring Paths and Semantic Similarity

Bridget McInnes
Ted Pedersen
Serguei Pakhomov

11/17/2009

Objective

Develop tools to automatically compute the semantic similarity between two concepts in the biomedical domain using measures originally developed for general English using the Unified Medical Language System (UMLS)

Motivation

- Clustering symptoms and disorders found in the text of clinical reports for post marking medication safety and surveillance
- Identification of patients for clinical studies
- Improving the sensitivity of document retrieval of scientific journals and clinical reports
- Development of terminologies and ontologies
- Clustering of biomedical documents
- Word sense disambiguation

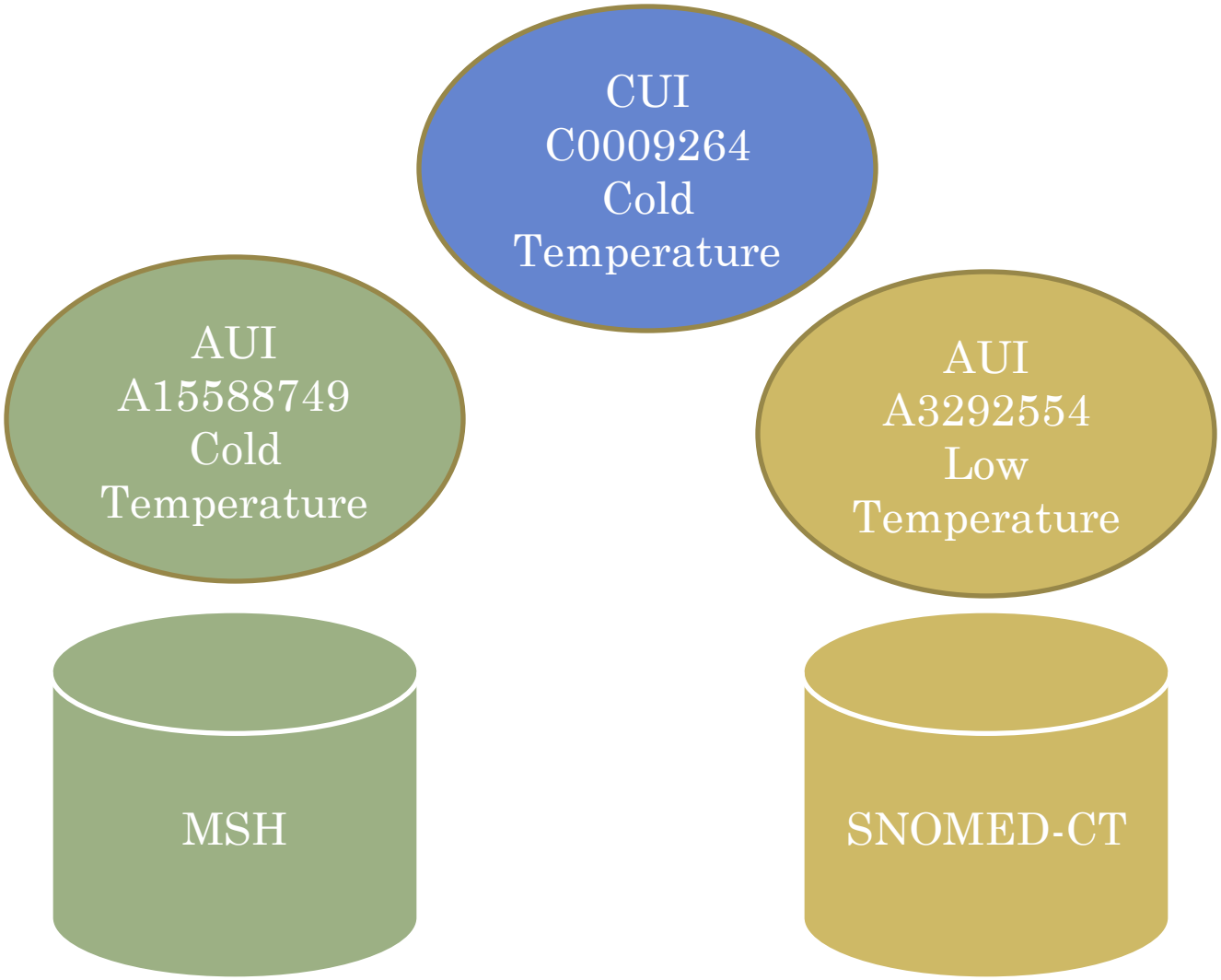
Unified Medical Language System

- Knowledge representation framework
- Contains 3 Main components:
  - **Metathesaurus**
  - Semantic Network
  - SPECIALIST Lexicon

Metathesaurus

- Semi-automatically integrates biomedical concepts from over a 100 controlled medical terminologies
- Source vocabularies are organized based on their Atomic Unique Identifiers
- Metathesaurus is organized based on their Concept Unique Identifier (CUI)

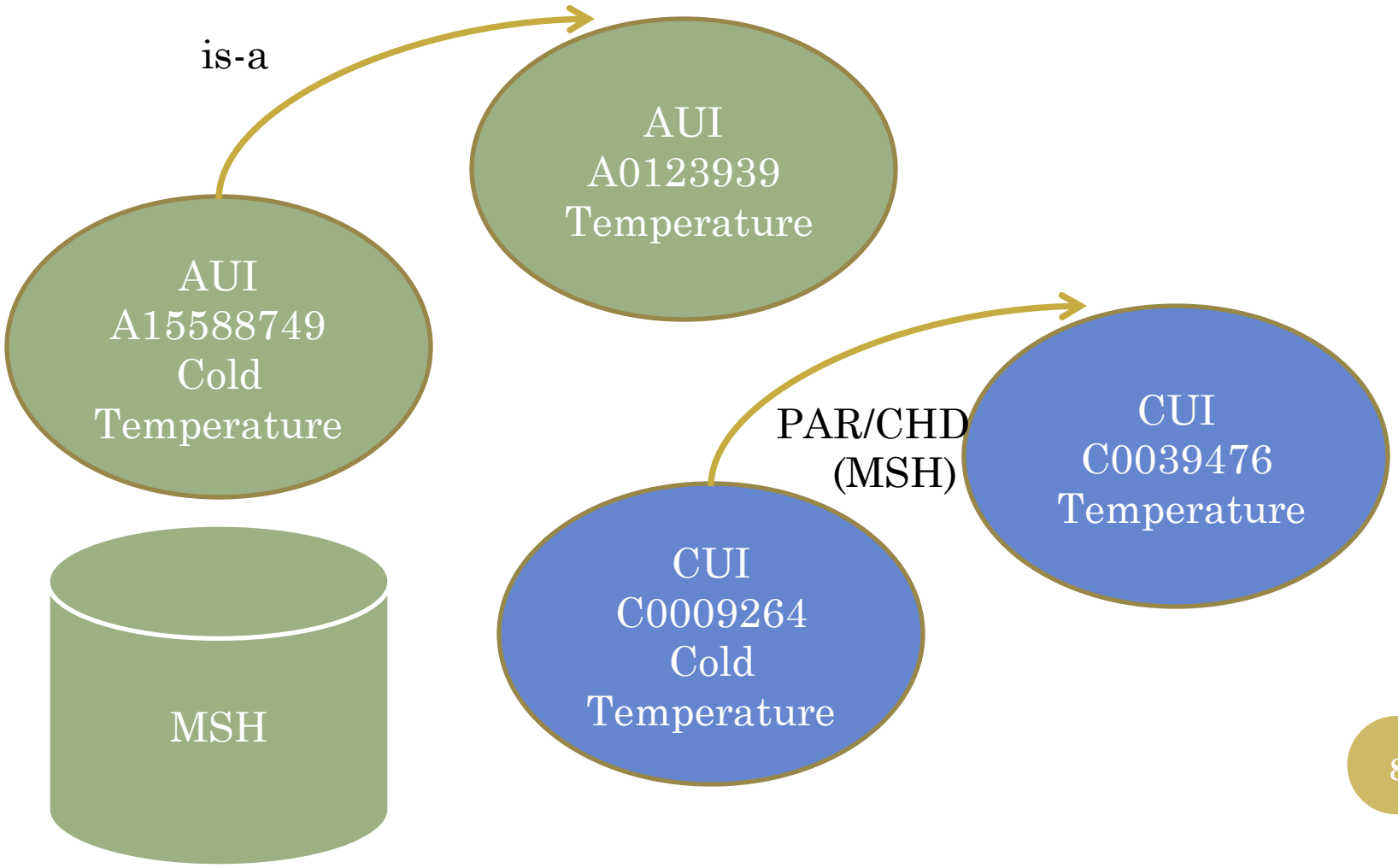
Concept Unique Identifiers (CUIs)



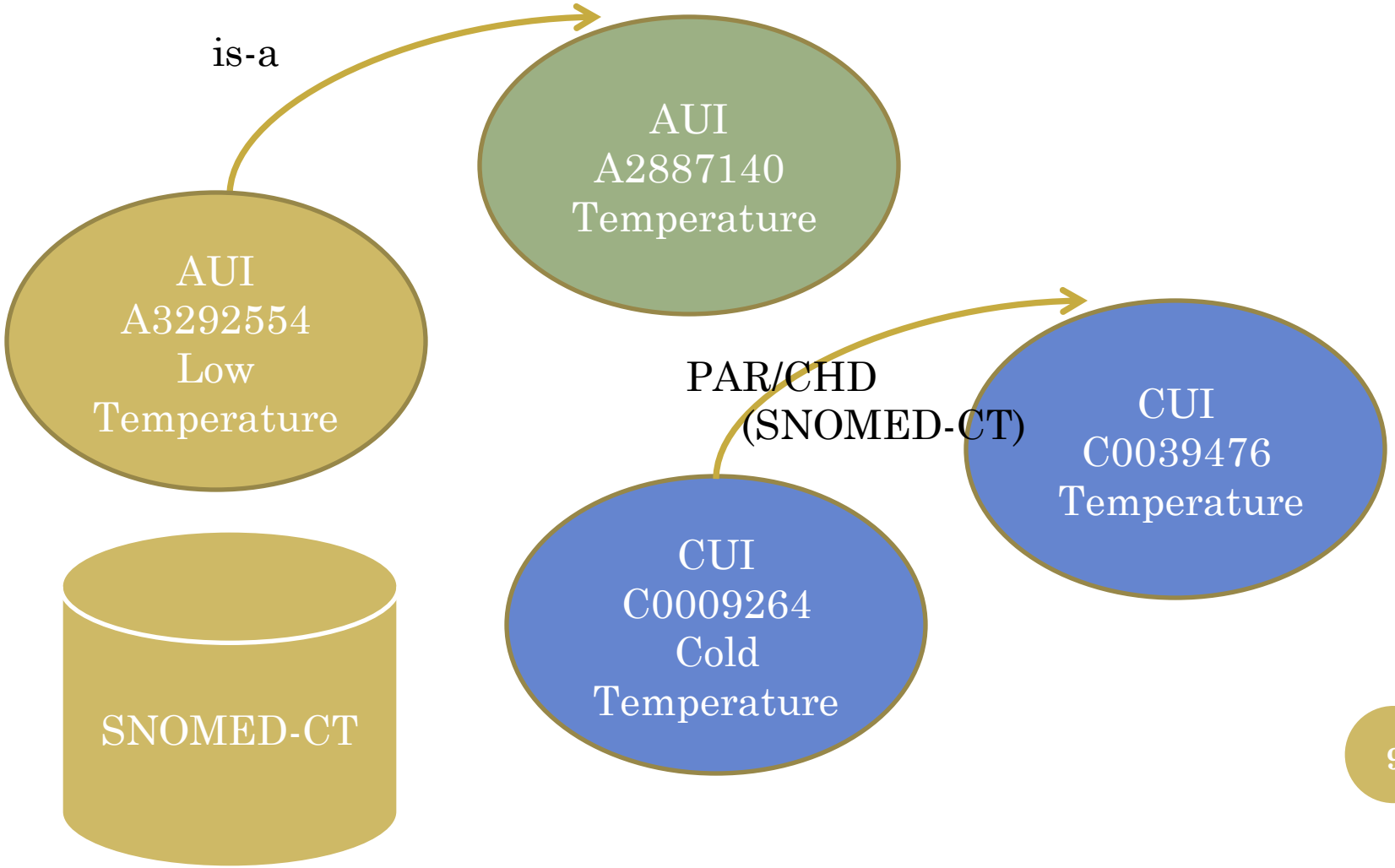
CUI Information

- The concepts (AUIs) from the source vocabularies may contain information about the concept such as its
  - Definition
  - Relation information between the concepts
- The information from the AUIs can be obtained through their respective CUIs

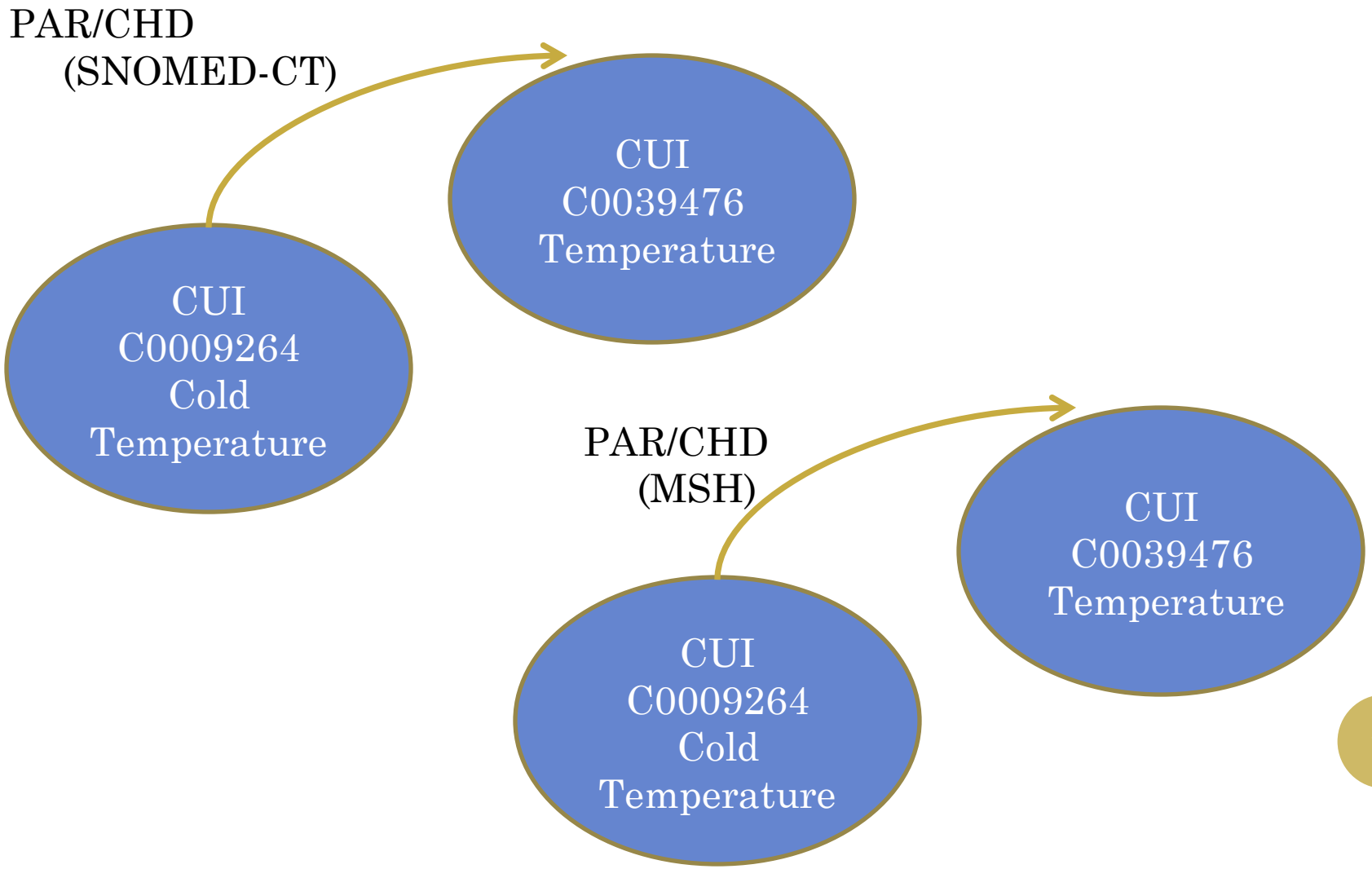
Relations between CUIs in MSH



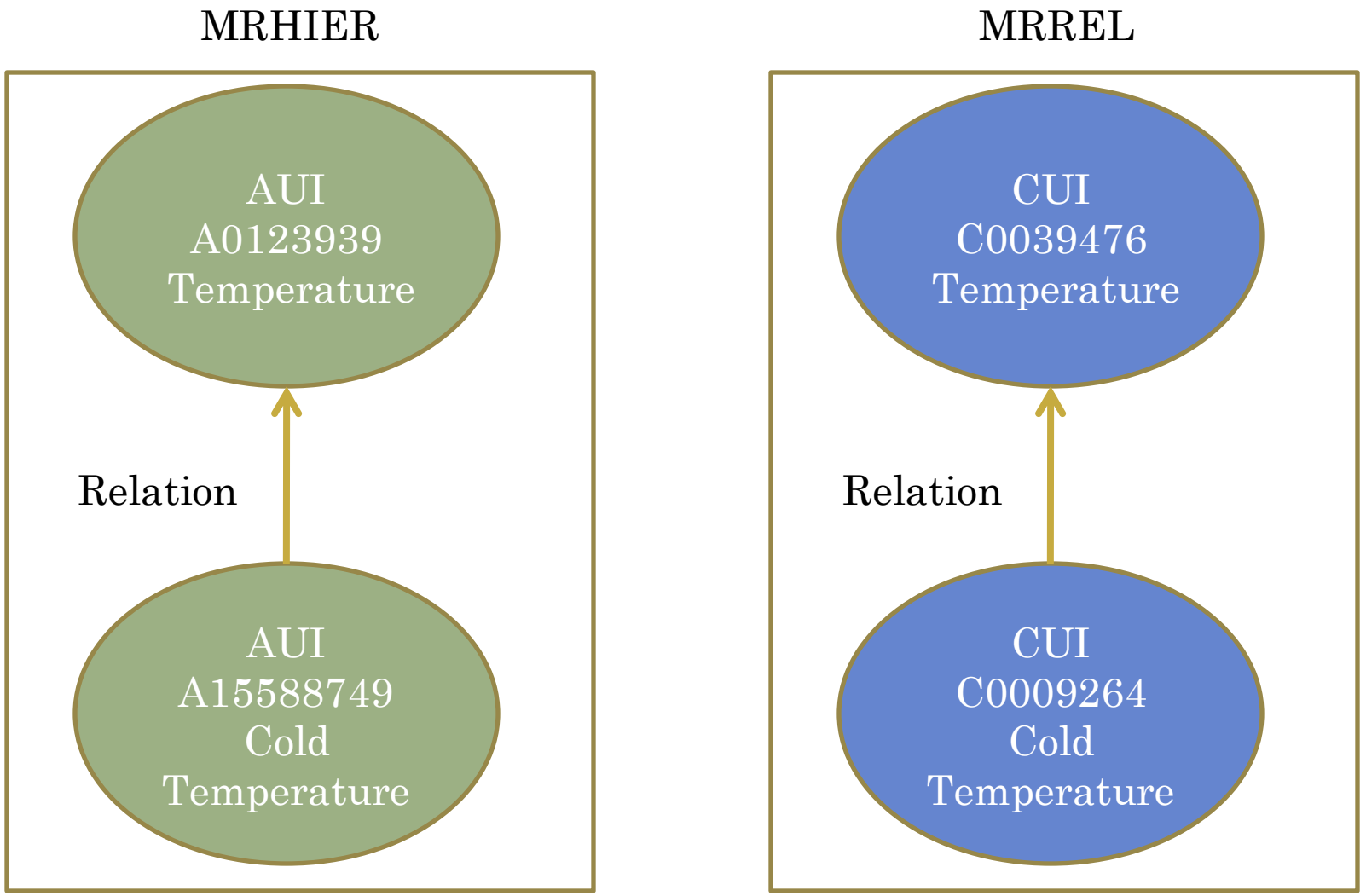
Relations between CUIs in SNOMED-CT



Multiple Relations



Relation Information



MRREL AND MRHIER

- MRHIER
  - Contains the full path to root relations between AUIs from each of the sources
    - is-a
    - part-of
- MRREL
  - Contains the pairwise relations between CUIs
  - Relations:
    - PAR/CHD
    - RB/RN
- It is possible to generate MRHEIR from MRREL except for the following sources:
  - AIR
  - MSH
  - SNM2
  - USPMG
  - OMS

CUI versus AUI Hierarchy

- The benefit of using CUIs
  - Ability to obtain the relation information between concepts across sources
  - Ability to obtain the relation information between concepts using more than one type of relation:
    - **PAR/CHD – parent/child (relation in MRHIER)**
    - RB/RN – narrower/broader
    - SIB – sibling
    - RL – concepts are similar or 'alike'

- The benefit of using AUIs
  - Ability to obtain relation information (PAR/CHD) between concepts in the same source very quickly
  - incorporates tree positional information for sources such as MSH

UMLS-Query by Shah and Musen, 2008

UMLS-Interface

- Perl interface to the UMLS present locally in a MySQL database.

- Its main purpose is to returns path information about CUIs using the relation information in MRREL
  - All possible paths to the root
  - Shortest path between two concepts

UMLS-Similarity

- A suite of perl modules that implement a number of path-based semantic similarity measures to determine the similarity between two CUIs in the UMLS
  - Measures are path-based because they rely on the location of the concepts in a hierarchy
  - The path information is obtained using UMLS-Interface
- Semantic Similarity Measures:
  - Path measure
  - Conceptual Distance (Rada, et. al, 1989)
  - Leacock and Chodorow, 1998
  - Wu and Palmer, 1994
  - Nguyen and Al-Mubaid, 2006

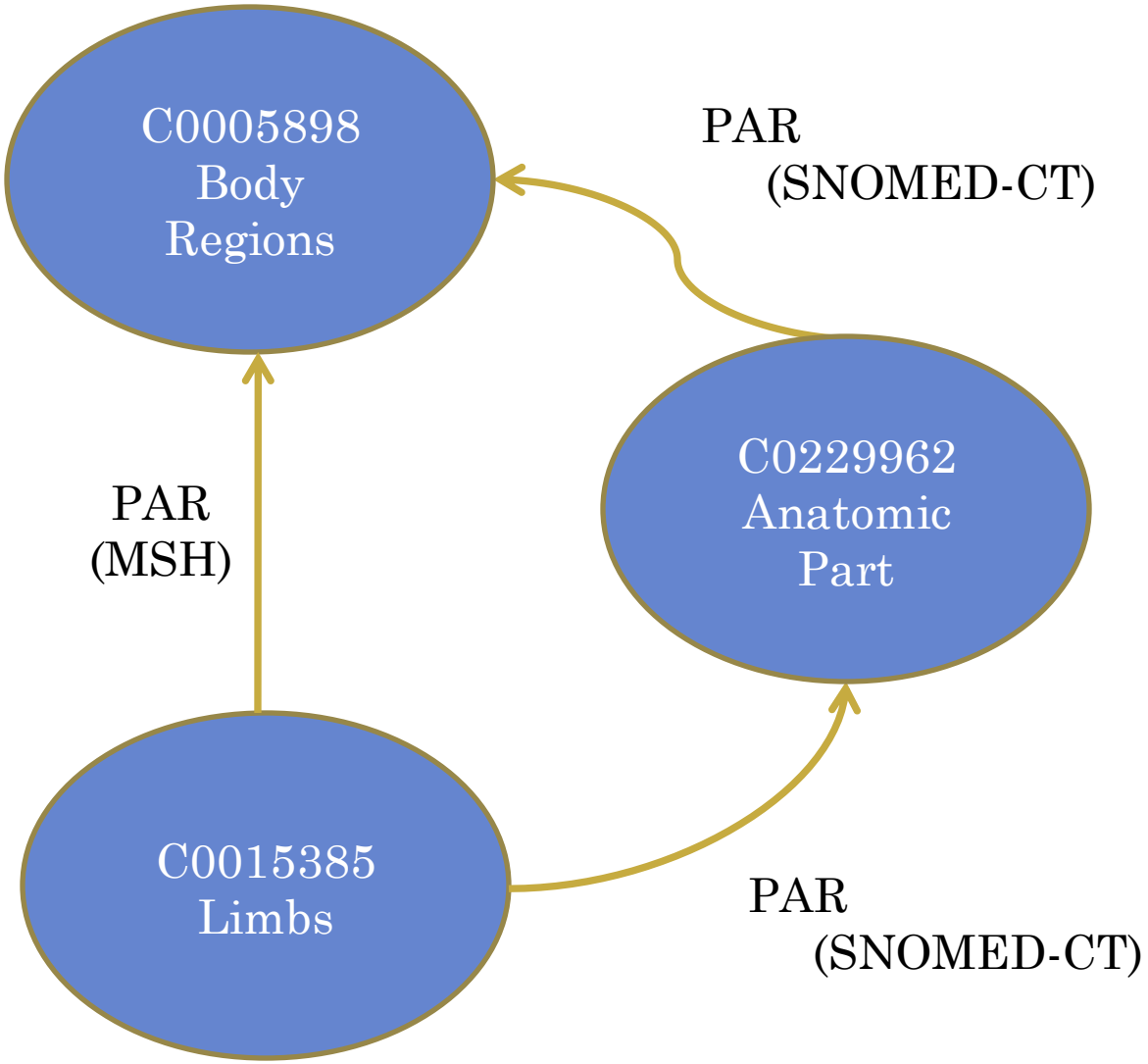
Semantic Similarity Example

- Path measure

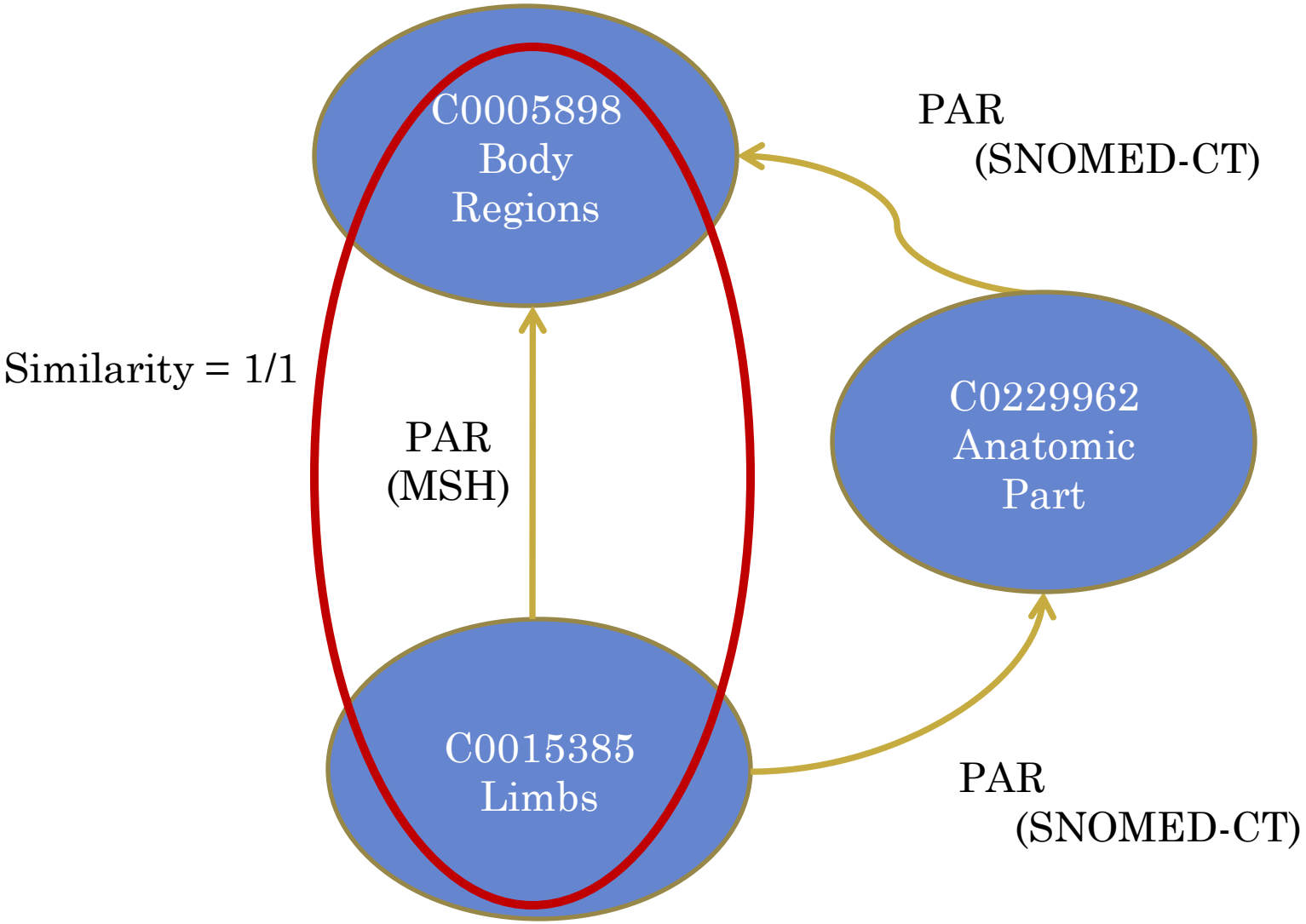
$$\text{Sim}(c1,c2) = \frac{1}{N}$$

where N = # links in the shortest path between the two concepts c1 and c2

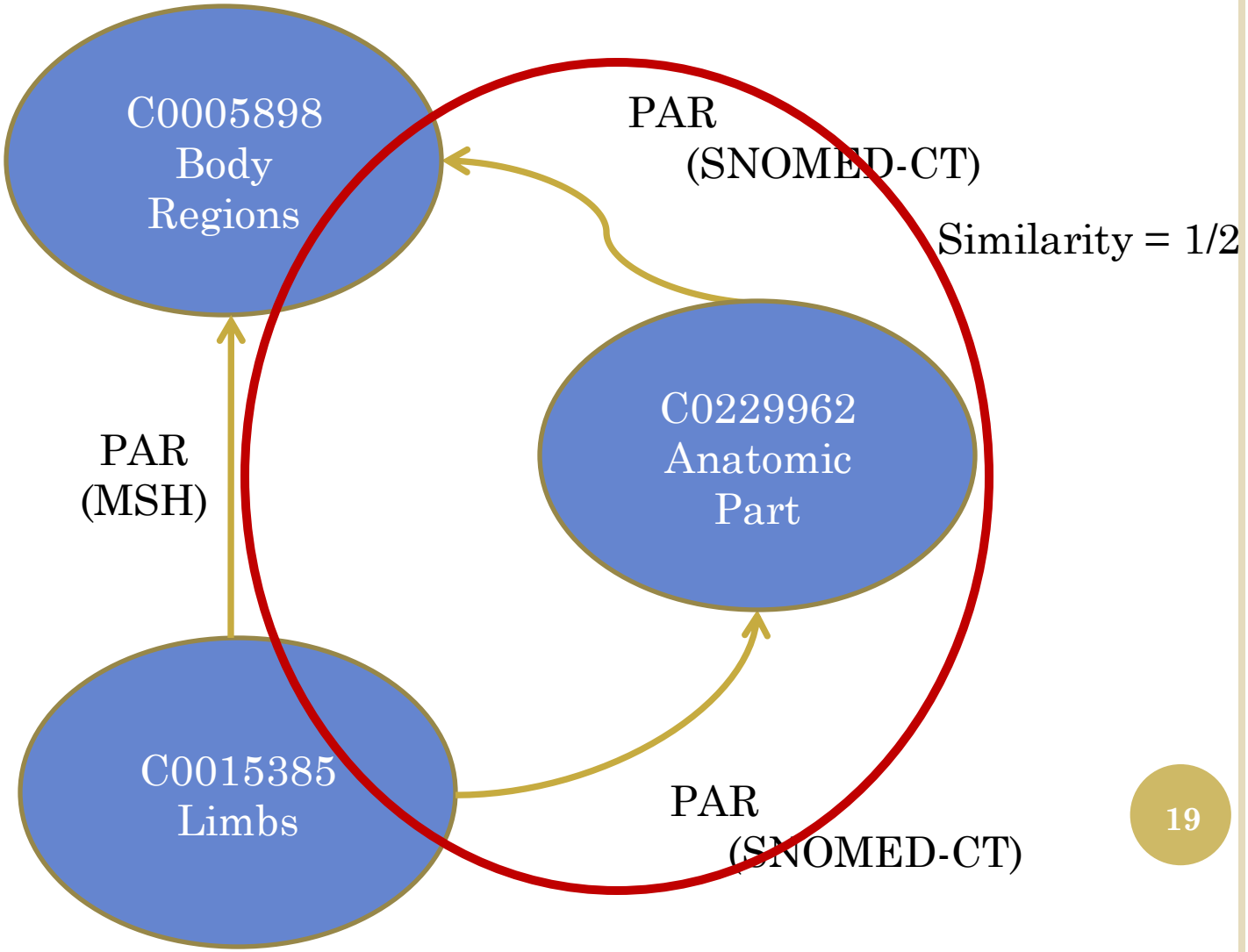
Similarity Given Specified Sources



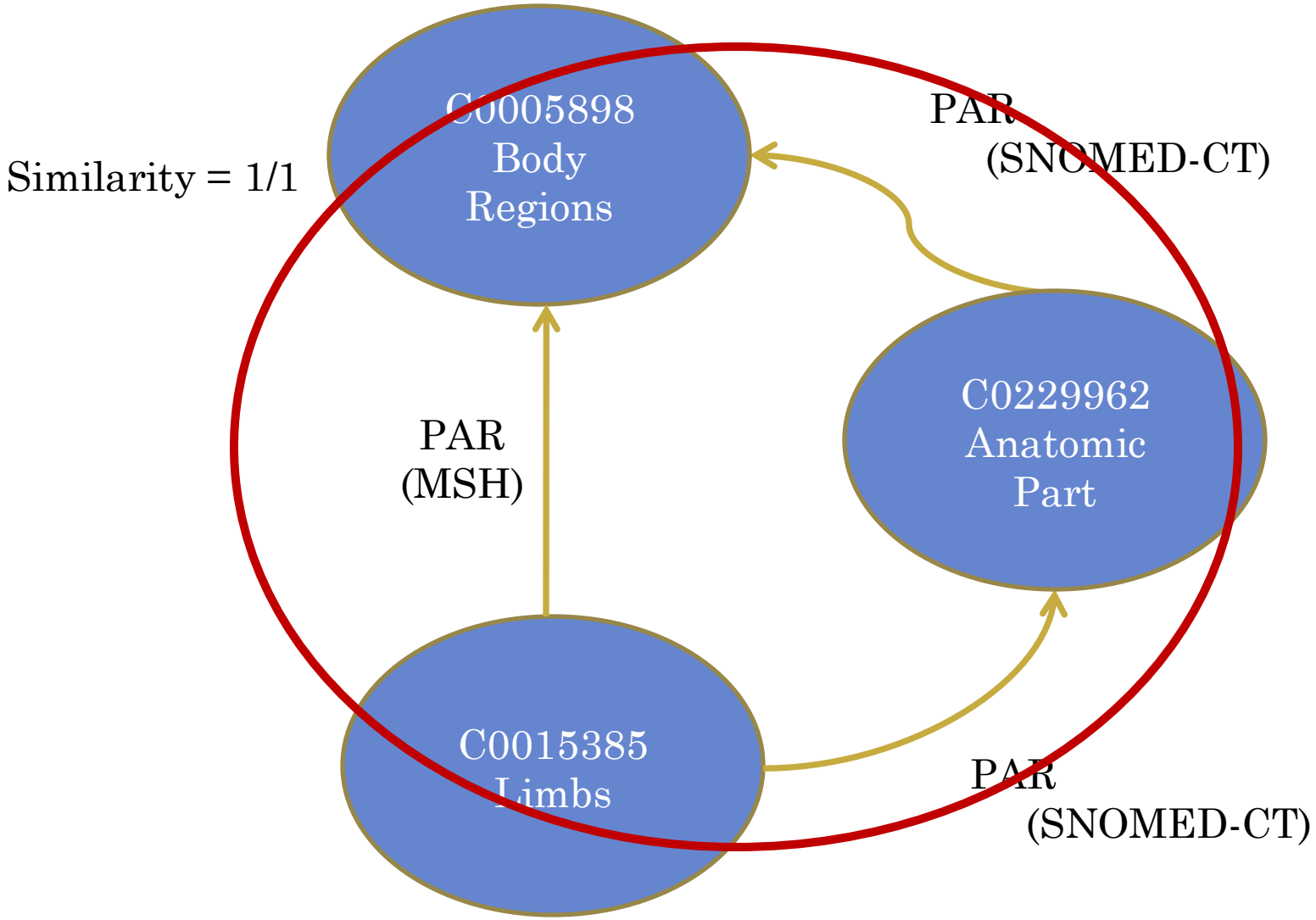
Similarity Given Specified Sources



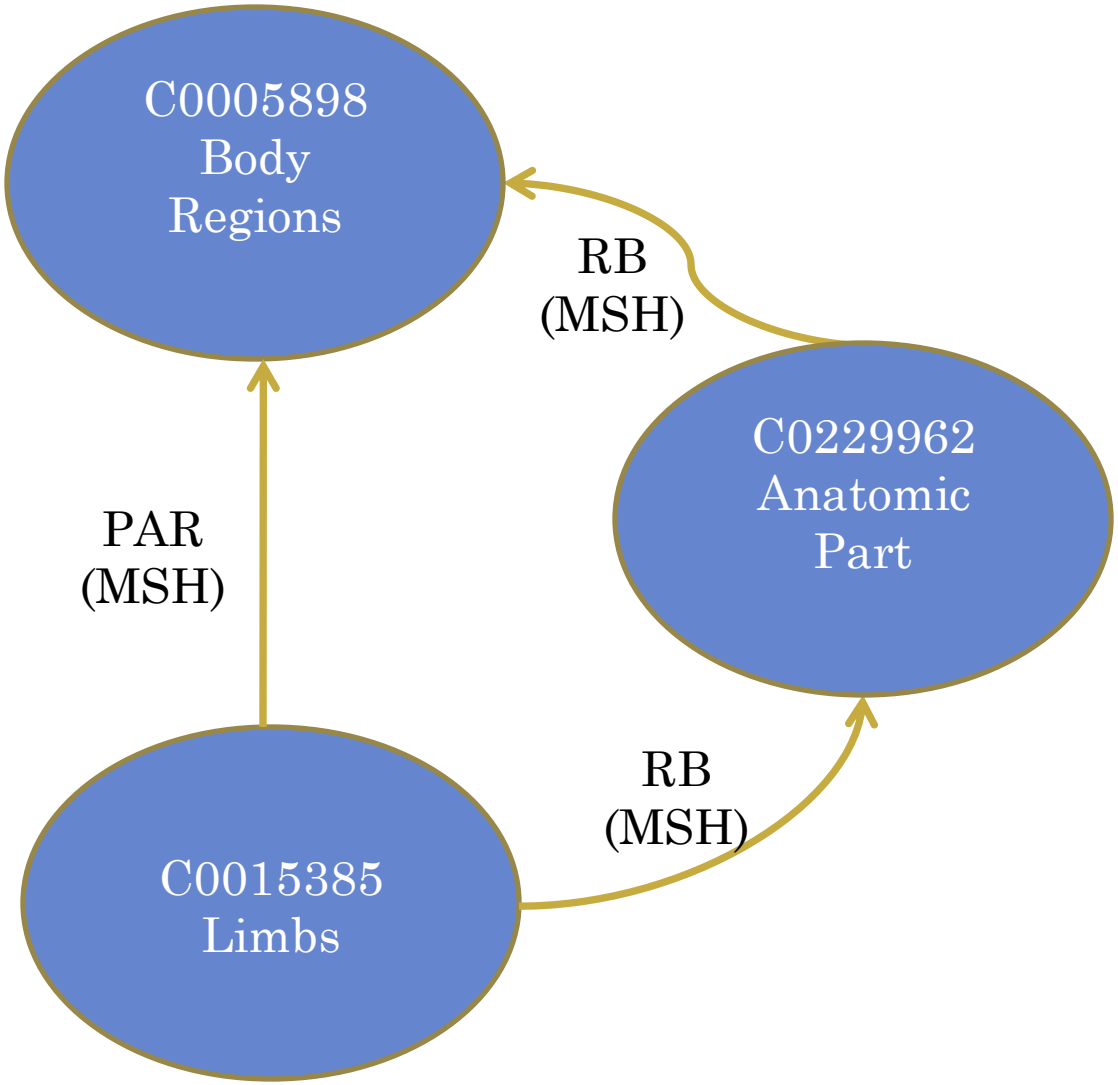
Similarity Given Specified Sources



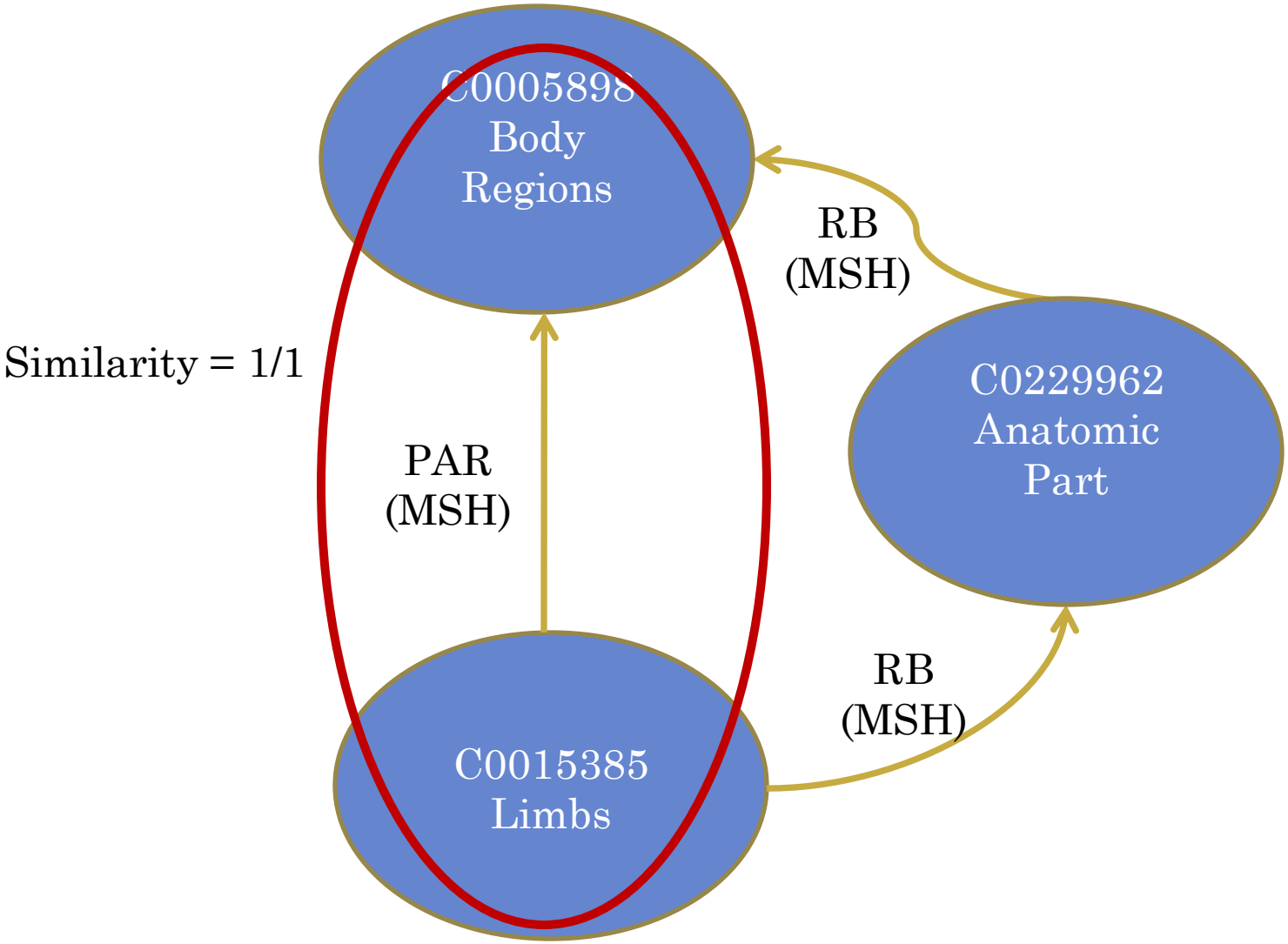
Similarity Given Specified Sources



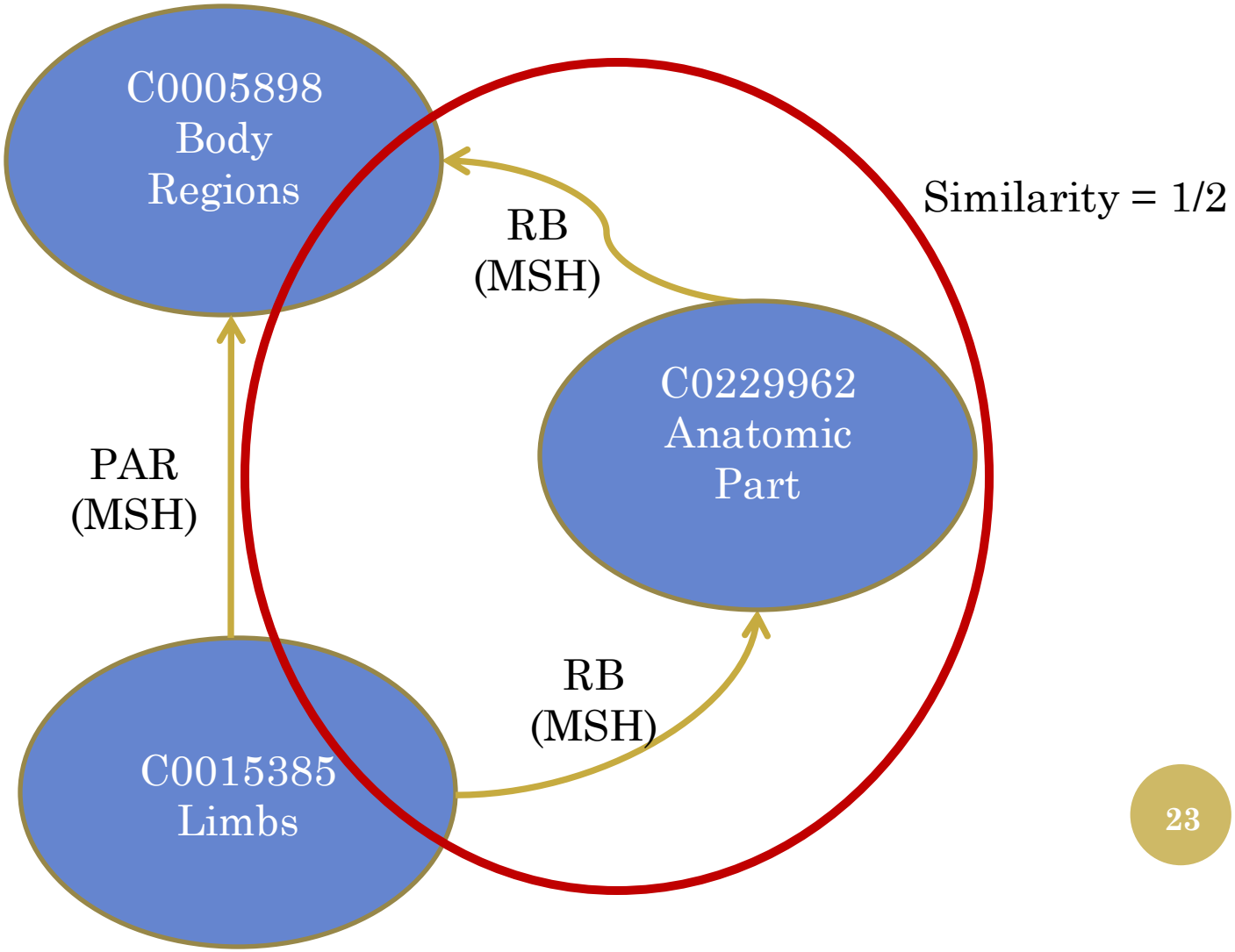
Similarity Given Specified Relations



Similarity Given Specified Relations

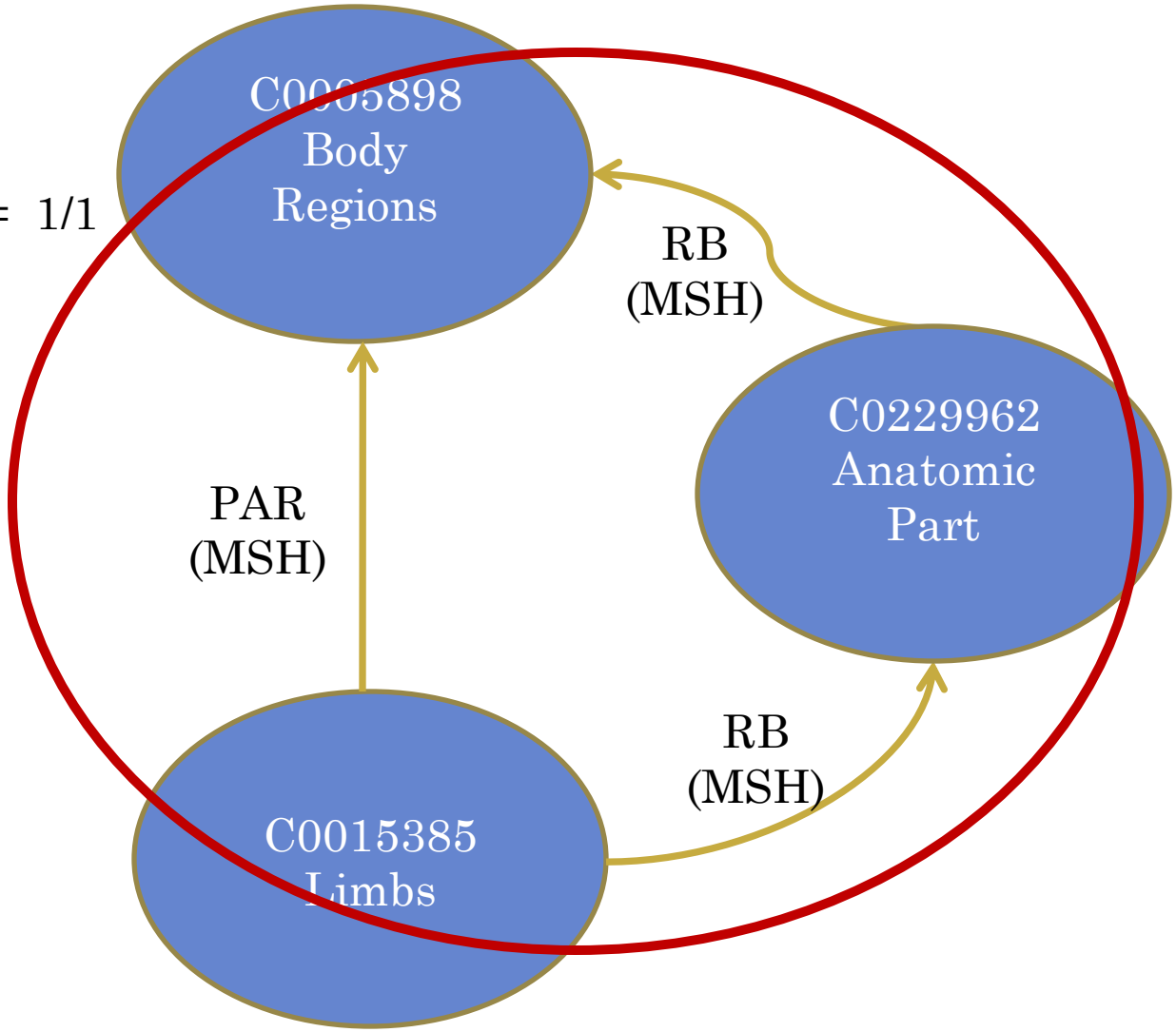


Similarity Given Specified Relations



Similarity Given Specified Relations

Similarity = 1/1



Functional Validation

- Comparison with Previous Work:
  - Pedersen, et al. 2007
  - Nguyen and Al-Mubaid, 2006
  - Caviedes and Cimino, 2004

Pedersen, et al.

- Semantic Similarity Measures
  - Path
  - Leacock and Chodorow, 1998
- Source
  - SNOMEDCT
- Data
  - 29 medical terms pairs
  - Similarity determined by:
    - 9 Medical Coders
    - 3 Physicians
  - 4 Point Scale
    - 4 – practically synonymous
    - 3 – related
    - 2 – marginally related
    - 1 - unrelated
- Spearman's Rank Correlation Coefficient

Comparison with Pedersen, et al.

- Semantic Similarity Measures
  - Path
  - Leacock and Chodorow, 1998
- Source: SNOMED-CT from UMLS 2008AB
- Relations: PAR/CHD
- Comparison with human annotations
  - Spearman Rank Correlation Coefficient

Comparison with Pedersen, et al.

Measure		Physician	Coder
path	**Pedersen, et. al.**	**0.36**	**0.51**
	UMLS-Similarity	**0.35**	**0.50**
Leacock and Chodorow	Pedersen, et. al.	0.35	0.50
	UMLS-Similarity	0.35	0.50

Comparison with Pedersen, et al.

Measure		Physician	Coder
path	Pedersen, et. al.	0.36	0.51
	UMLS-Similarity	0.35	0.50
Leacock and Chodorow	**Pedersen, et. al.**	**0.35**	**0.50**
	UMLS-Similarity	**0.35**	**0.50**

Nguyen and Al-Mubaid

- Semantic Similarity Measures
  - Nguyen and Al-Mubaid, 2006
  - Leacock and Chodorow, 1998
  - Wu and Palmer, 1994
  - Path
- Source: MSH
- Same Dataset created by Pedersen, et al.
  - Data
    - 29 medical terms pairs
    - Similarity determined by:
      - 9 Medical Coders
      - 3 Physicians
- Spearman's Rank Correlation Coefficient

Comparison with Nguyen and Al-Mubaid

- Semantic Similarity Measures
  - Nguyen and Al-Mubaid, 2006
  - Leacock and Chodorow, 1998
  - Wu and Palmer, 1994
  - Path
- Source: MSH from UMLS 2008AB
- Relations: PAR/CHD
- Comparison with human annotations
  - Spearman Rank Correlation Coefficient

Comparison with Nguyen and Al-Mubaid

Measure		Physician	Coder
path	Nguyen and Al-Mubaid	0.63	0.85
	UMLS-Similarity	0.49	0.58
Leacock and Chodorow	Nguyen and Al-Mubaid	0.67	0.86
	UMLS-Similarity	0.49	0.58
Wu and Palmer	Nguyen and Al-Mubaid	0.65	0.79
	UMLS-Similarity	0.45	0.54
Nguyen and Al-Mubaid	Nguyen and Al-Mubaid	0.67	0.86
	UMLS-Similarity	0.45	0.55

Caviedes and Cimino

- Semantic Similarity Measure
  - Conceptual Distance – Rada, et al.
- Source: MSH
- Relations: PAR/CHD
- Data
  - 10 medical terms pairs using following CUIs
    - Digestive system disease: C0012242
    - Peptic esophagitis: C0014869
    - Psychotherapy: C0033968
    - Thirst: C0039971
    - Thoracic duct: C0039979

Comparison with Caviedes and Cimino

- Semantic Similarity Measures
  - Conceptual Distance
    - Originally proposed by Rada, et. al., 1989
- Source: MSH from UMLS 2008AB
  - Relations: PAR/CHD
- Comparison between the Conceptual Distance Scores

Comparison with Caviedes and Cimino

CUI Pairs	Caviedes and Cimino	UMLS-Similarity
C0012242-C0014869	**3**	**3**
C0012242-C0033968	**5**	**5**
C0033968-C0039971	**6**	**6**
C0012242-C0039971	**7**	**7**
C0012242-C0039979	7	6
C0033968-C0039979	8	9
C0014869-C0033968	**8**	**8**
C0014869-C0039971	**10**	**10**
C0014869-C0039979	10	11
C0039971-C0039979	10	11

Comparison with Caviedes and Cimino

CUI Pairs	Caviedes and Cimino	UMLS-Similarity
C0012242-C0014869	3	3
C0012242-C0033968	5	5
C0033968-C0039971	6	6
C0012242-C0039971	7	7
C0012242-C0039979	**7**	**6**
C0033968-C0039979	**8**	**9**
C0014869-C0033968	8	8
C0014869-C0039971	10	10
C0014869-C0039979	**10**	**11**
C0039971-C0039979	**10**	**11**

Results

- The results show that UMLS-Similarity can be used to reproduce the results reported by:
  - Pedersen, et al.
  - Caviedes and Cimino

Results

- The correlation results obtained by UMLS-Similarity and reported by Nguyen and Al-Mubaid vary
  - Different versions of MSH were used to conduct the experiment
  - Possibly different mappings of the terms to CUIs in MSH were used
  - Information used by Nguyen and Al-Mubaid comes directly from MSH which is located in MRHEIR and as PAR/CHD relations in MRREL
    - It is not possible to generate MRHIER from MRREL because the full path-to-root is a transitive closure of the pairwise PAR/CHD relations which does not hold true for MSH because a MSH concept may have different children depending on its tree position

Conclusions

- UMLS-Similarity
  - Used to determine the similarity between two concepts given a specified set of sources and relations
  - Contains the following similarity measures
    - Path measure
    - Conceptual Distance proposed Rada, et. al. 1989
    - Leacock and Chodorow, 1998
    - Wu and Palmer, 1994
    - Nguyen and Al-Mubaid, 2006
- UMLS-Interface
  - Used to obtain path information about a CUI given a specified set of sources and relations

Future Work

- UMLS-Interface
  - Improve the efficiency in which the path information is stored
- UMLS-Similarity
  - Information Content Similarity Measures
    - Resnik, 1995
    - Jiang and Conrath, 1997
    - Lin, 1997
  - Relatedness Measures
    - Patwardhan, 2003

Take Home Message #1

UMLS-Interface can used to extract path information about a concept given a specified set of sources and relations.

Take Home Message #2

UMLS-Similarity can be used to compute the semantic similarity between two concepts given a specified set of sources and relations.

Availability

- UMLS-Interface
  - http://search.cpan.org/dist/UMLS-Interface
- UMLS-Similarity
  - http://search.cpan.org/dist/UMLS-Similarity

Thank You

- We would like to thank
  - Kin Wah Fung,
  - Olivier Bodenreider,
  - Jan Willis
  - Lan Aronson

- The research was supported in parts by:
  - Fellowships:
    - NLM Research Participation Program
    - GAANN fellowship from US Dept. of Ed.
  - Grants
    - IR01LM009623-01A2 from NIH, NLM