

Measuring the Similarity and Relatedness of Concepts in the Medical Domain : IHI 2012 Tutorial

Ted Pedersen, Ph.D. *
Serguei Pakhomov, Ph.D. #
Bridget McInnes, Ph.D. #
Ying Liu, Ph.D. #

University of Minnesota

* Department of Computer Science, Duluth

College of Pharmacy, Twin Cities

{tpederse,pakh0002,bthomson,liux0395}@umn.edu

Acknowledgment

- The development of this tutorial and the work that underlies it was supported in part by grant 1R01LM009623-01A2 from the National Library of Medicine, National Institutes of Health.
- The contents of this tutorial are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

What (we hope) you will learn!

- The distinction between semantic similarity and relatedness (and why both are useful)
- How to measure using information from ontologies, definitions, and corpora
- How to use the freely available software UMLS::Similarity and UMLS::Interface
- How to conduct experiments using freely available reference standards
- How to integrate these measures into clinical NLP applications

Outline

- Introduction to the measures
 - Pedersen, 30 minutes
- Using path and information content measures
 - McInnes, 45 minutes
- Using vector and lesk measures
 - Liu, 15 minutes
- Evaluating measures and deploying
 - Pakhomov, 30 minutes

Logistics

- Questions? Just ask!
 - We've planned for ~5 minutes of questions each half hour, but if yours are more extensive or specific to your situation please consider asking after tutorial or via email
- Mailing list, software, data, web interfaces, TUTORIAL SLIDES, and more information :
 - <http://umls-similarity.sourceforge.net>
- Need a break? Feel free, but on your own (and be quick about it! ;)
 - we'll keep going til end of session (5:00 pm)

Introducing Measures of Semantic Similarity and Relatedness (without tears)

Ted Pedersen
Department of Computer Science
University of Minnesota
Duluth, MN 55812
tpederse@d.umn.edu
<http://www.d.umn.edu/~tpederse>

What are we measuring?

- Concept pairs
 - Assign a numeric value that quantifies how similar or related two concepts are
- Not words
 - Must know concept underlying a word form
 - Cold may be *temperature* or *illness*
 - Concept Mapping
 - Word Sense Disambiguation
 - This tutorial assumes that's been resolved

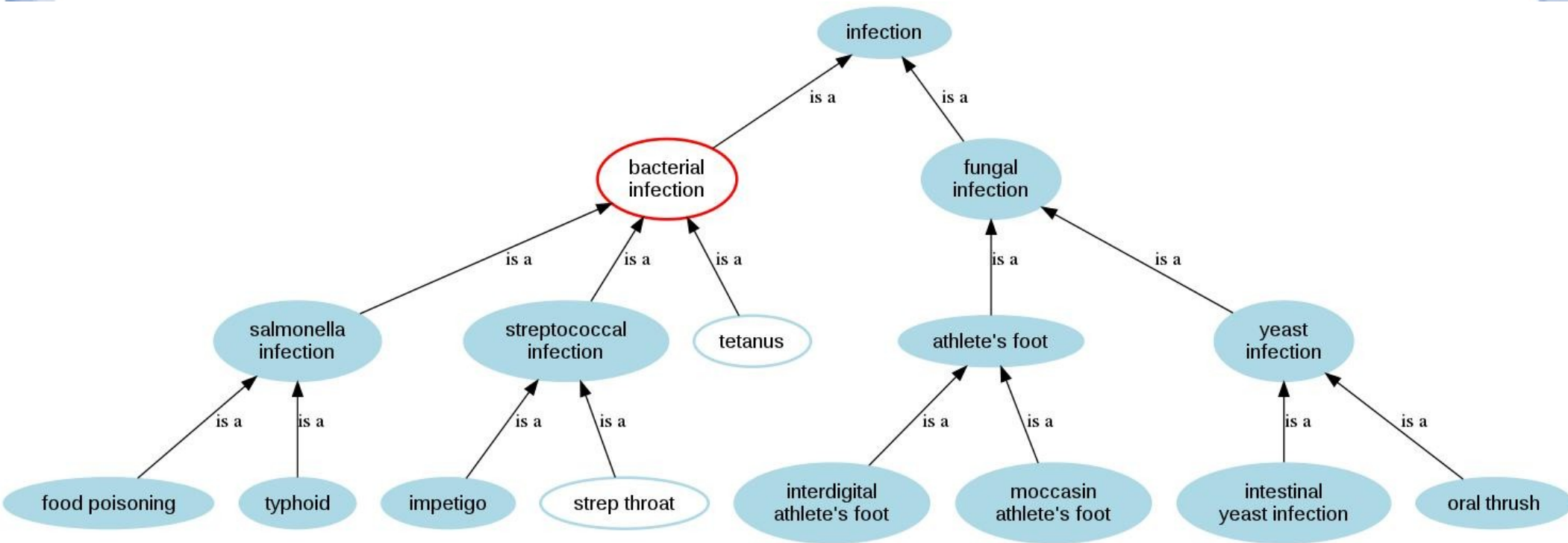
Why?

- Being able to organize concepts by their similarity or relatedness to each other is a fundamental operation in the human mind, and in many problems in Natural Language Processing and Artificial Intelligence
- If we know a lot about X, and if we know Y is similar to X, then a lot of what we know about X may apply to Y
 - Use X to explain or categorize Y

Similar or Related?

- Similarity based on is-a relations
 - How much is X like Y?
 - Share ancestor in is-a hierarchy
 - LCS : least common subsumer
 - Closer / deeper the ancestor the more similar
- *Tetanus* and *strep throat* are similar
 - both are kinds-of bacterial infections

Least Common Subsumer (LCS)



Similar or Related?

- Relatedness more general
 - How much is X related to Y?
 - Many ways to be related
 - is-a, part-of, treats, affects, symptom-of, ...
- *Tetanus* and *deep cuts* are related but they really aren't similar
 - (deep cuts can cause tetanus)
- All similar concepts are related, but not all related concepts are similar

Measures of Similarity

(all available in UMLS::Similarity)

- Path Based
 - **Rada et al., 1989 (path)**
 - Caviedes & Cimino, 2004 (cdist)
- Path + Depth
 - **Wu & Palmer, 1994 (wup)**
 - Leacock & Chodorow, 1998 (lch)
 - Zhong et al., 2002 (zhong)
 - Nguyen & Al-Mubaid, 2006 (nam)

Measures of Similarity

(all available in UMLS::Similarity)

- Path + Information Content
 - Resnik, 1995 (res)
 - Jiang & Conrath, 1997 (jcn)
 - **Lin, 1998 (lin)**

Path Based Measures

- Distance between concepts (nodes) in tree intuitively appealing
- Spatial orientation, good for networks or maps but not is-a hierarchies
 - Reasonable approximation sometimes
 - Assumes all paths have same “weight”
 - But, more specific (deeper) paths tend to travel less semantic distance
- Shortest path a good start, but needs corrections

Shortest is-a Path

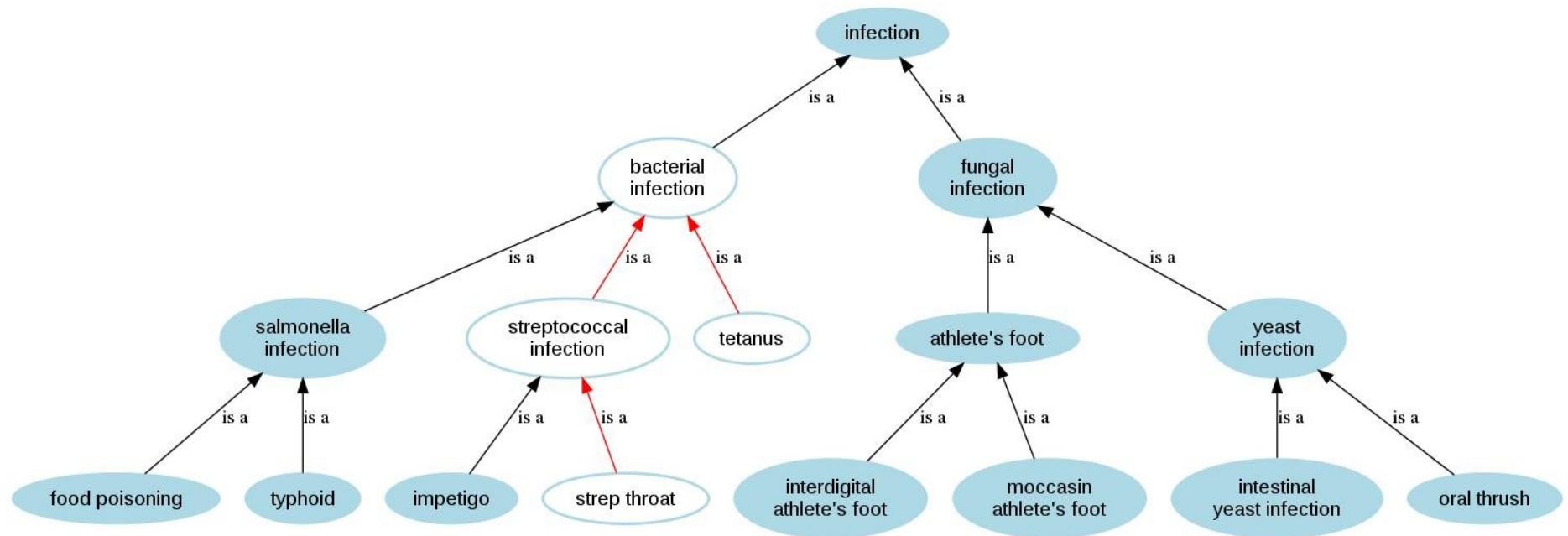
1

- $\text{path}(a,b) = \text{shortest is-a path}(a,b)$

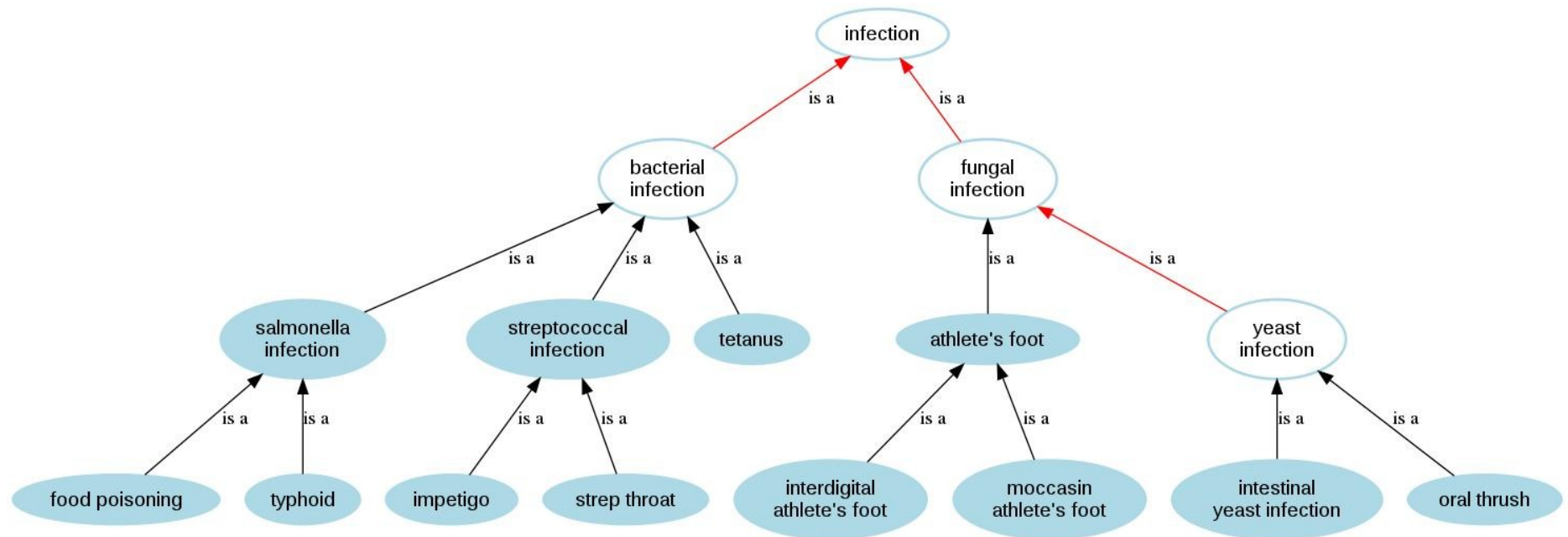
We count nodes...

- Maximum = 1
 - self similarity
 - $\text{path}(\text{tetanus}, \text{tetanus}) = 1$
- Minimum = $1 / (\text{longest path in isa tree})$
 - $\text{path}(\text{typhoid}, \text{oral thrush}) = 1/7$
 - $\text{path}(\text{food poisoning}, \text{strep throat}) = 1/7$
 - etc...

$path(strep\ throat, tetanus) = .25$



path (bacterial infections, yeast infections) = .25



?

- Are *bacterial infections* and *yeast infections* similar to the same degree as are *tetanus* and *strep throat* ?
- The path measure says “yes, they are.”

Path + Depth

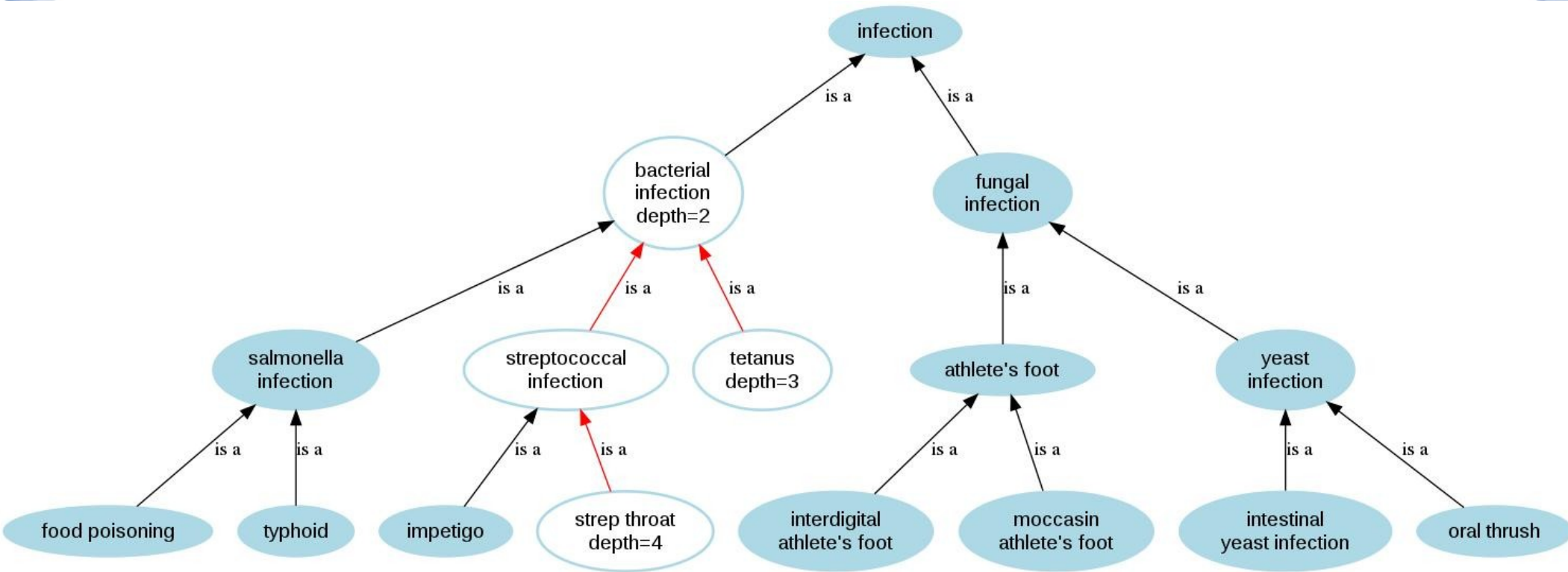
- Path only doesn't account for specificity
- Deeper concepts more specific
- Paths between deeper concepts travel less semantic distance

Wu and Palmer, 1994

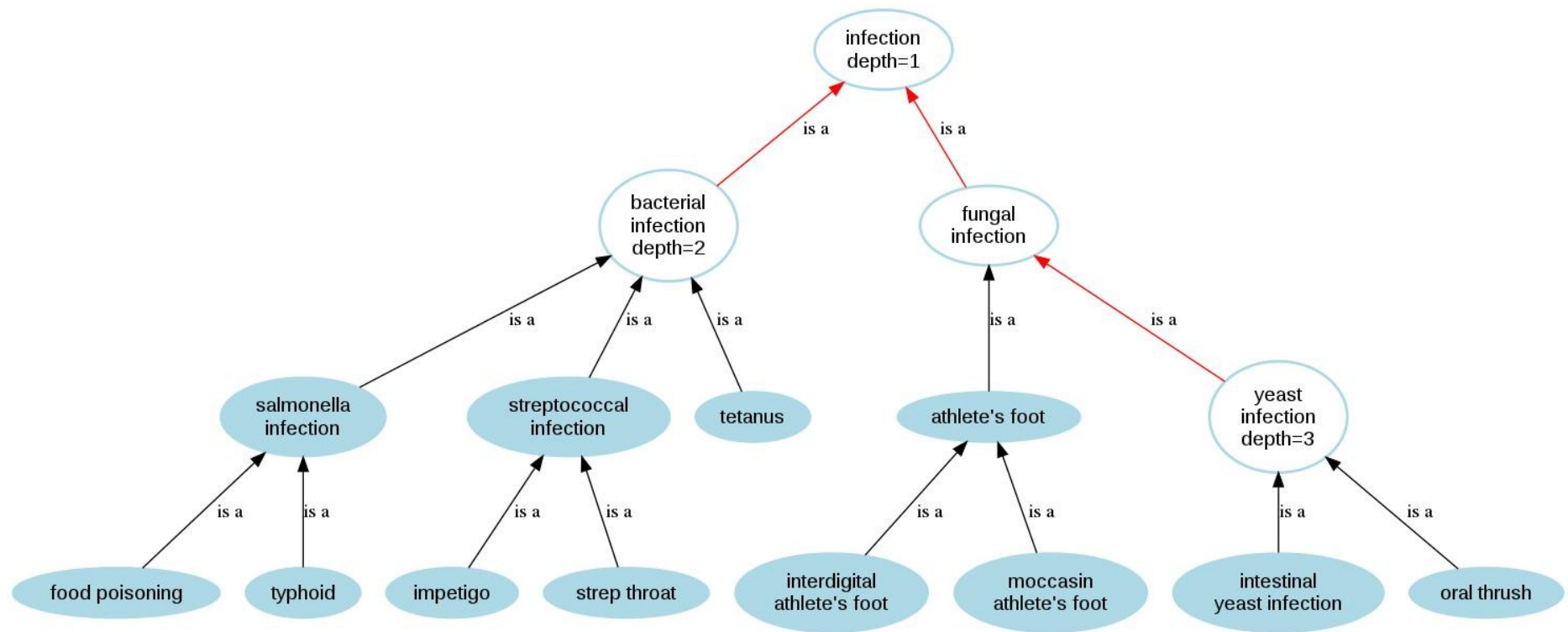
$$2 * \text{depth}(\text{LCS}(a,b))$$

- $\text{wup}(a,b) = \frac{\text{depth}(\text{LCS}(a,b))}{\text{depth}(a) + \text{depth}(b)}$
- $\text{depth}(x) = \text{shortest is-a path}(\text{root}, x)$

$$wup(\text{strep throat}, \text{tetanus}) = (2*2)/(4+3) = .57$$



$wup(\text{bacterial infections}, \text{yeast infections}) = (2 \cdot 1) / (2 + 3) = .4$



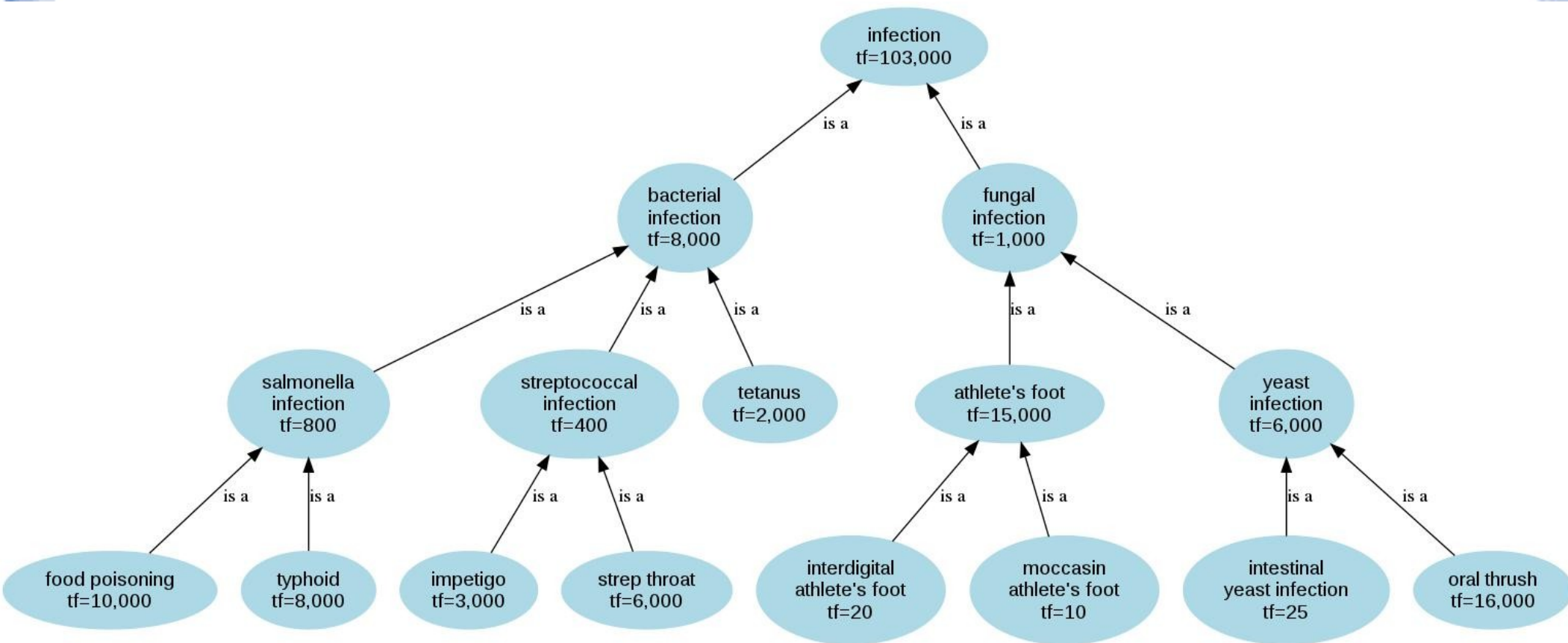
?

- Wu and Palmer say that *strep throat* and *tetanus* (.57) are more similar than are *bacterial infections* and *yeast infections* (.4)
- Path says that *strep throat* and *tetanus* (.25) are equally similar as are *bacterial infections* and *yeast infections* (.25)

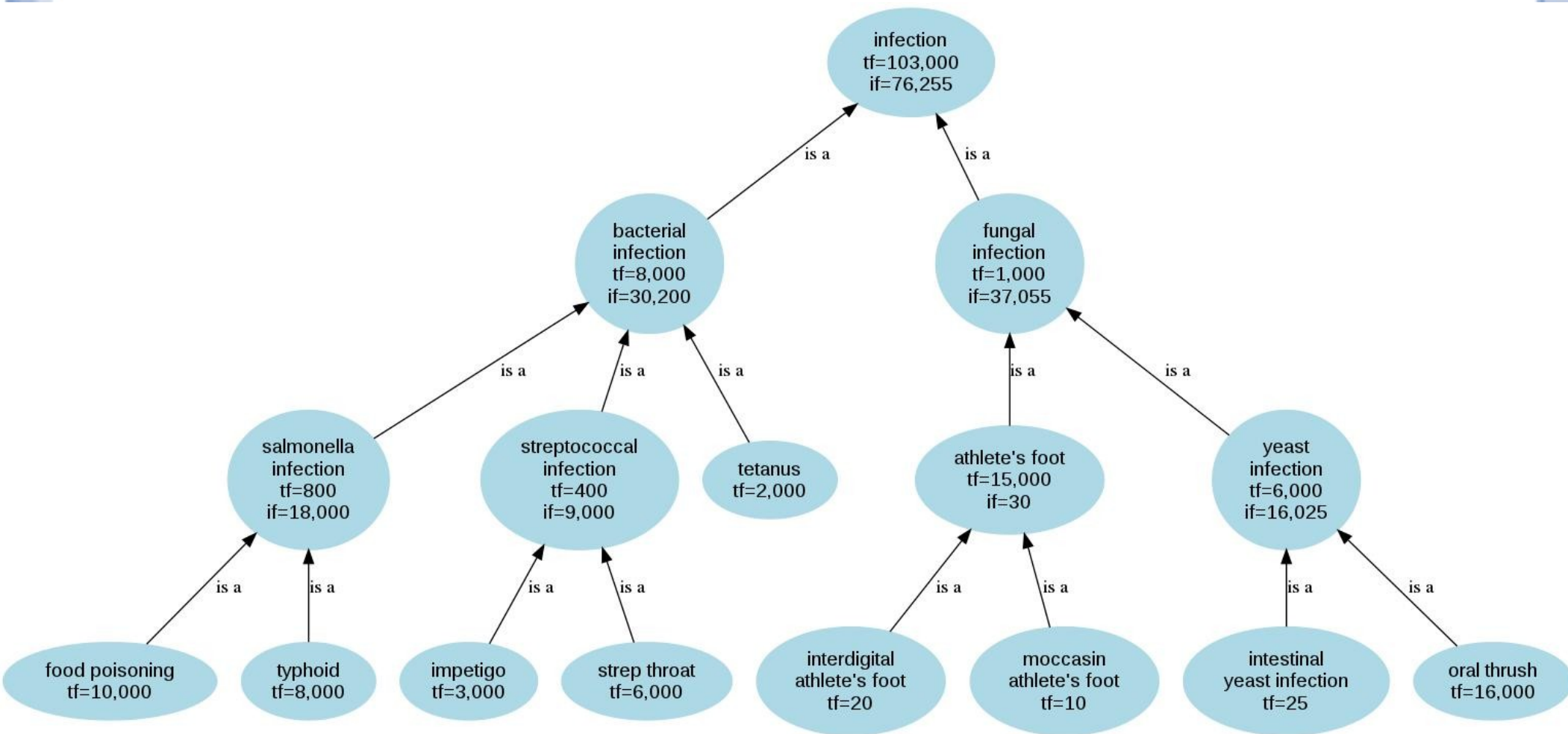
Information Content

- $ic(\text{concept}) = -\log p(\text{concept})$ [Resnik 1995]
 - Need to count concepts
 - Term frequency + Inherited frequency
 - $p(\text{concept}) = (tf + if) / N$
- Depth shows specificity but not frequency
- Low frequency concepts often much more specific than high frequency ones

Information Content term frequency (tf)

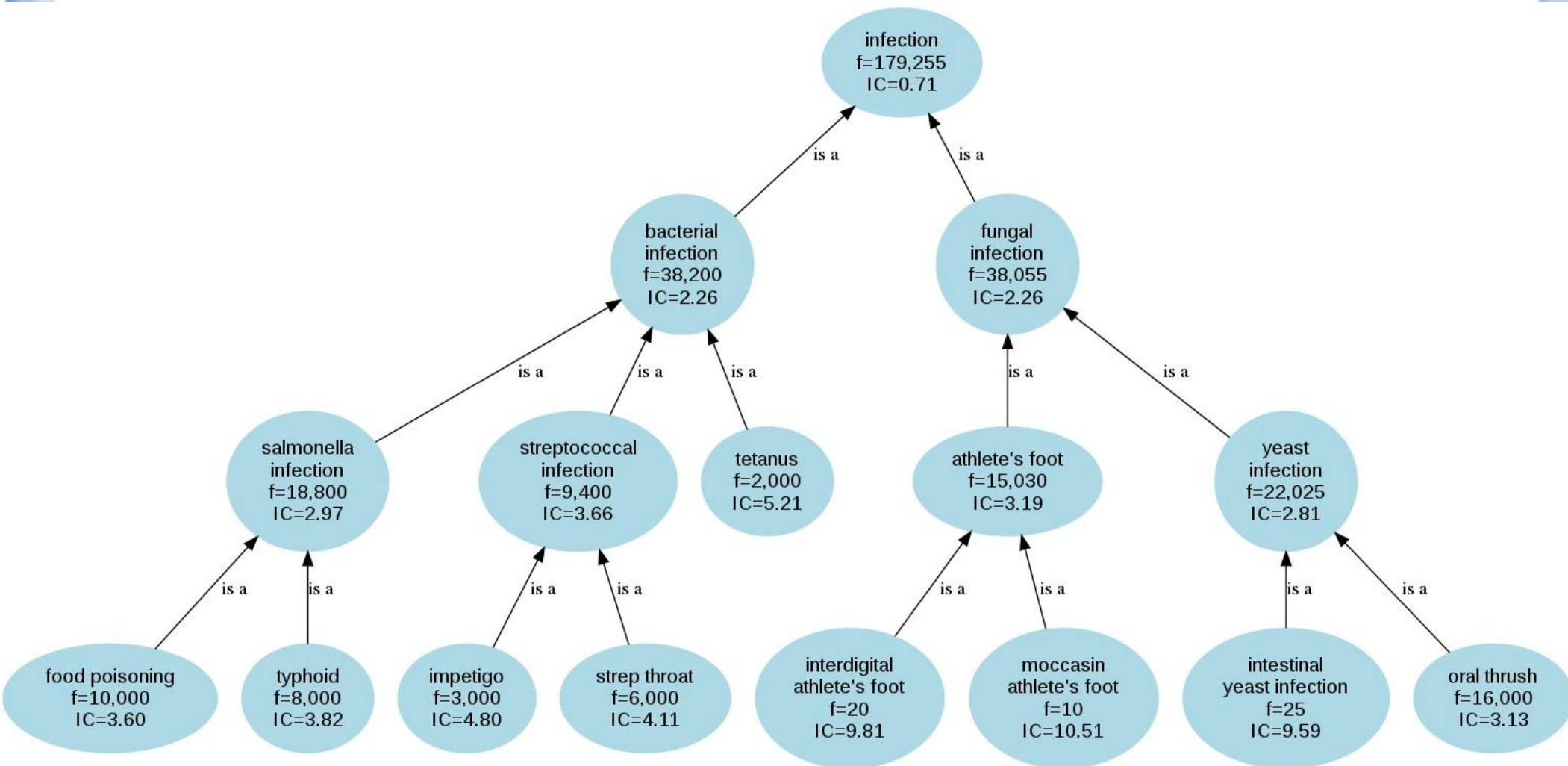


Information Content inherited frequency (if)



Information Content ($IC = -\log(f/N)$)

final count ($f = t_f + i_f$, $N = 365,820$)



Lin, 1998

$$2 * IC (LCS (a,b))$$

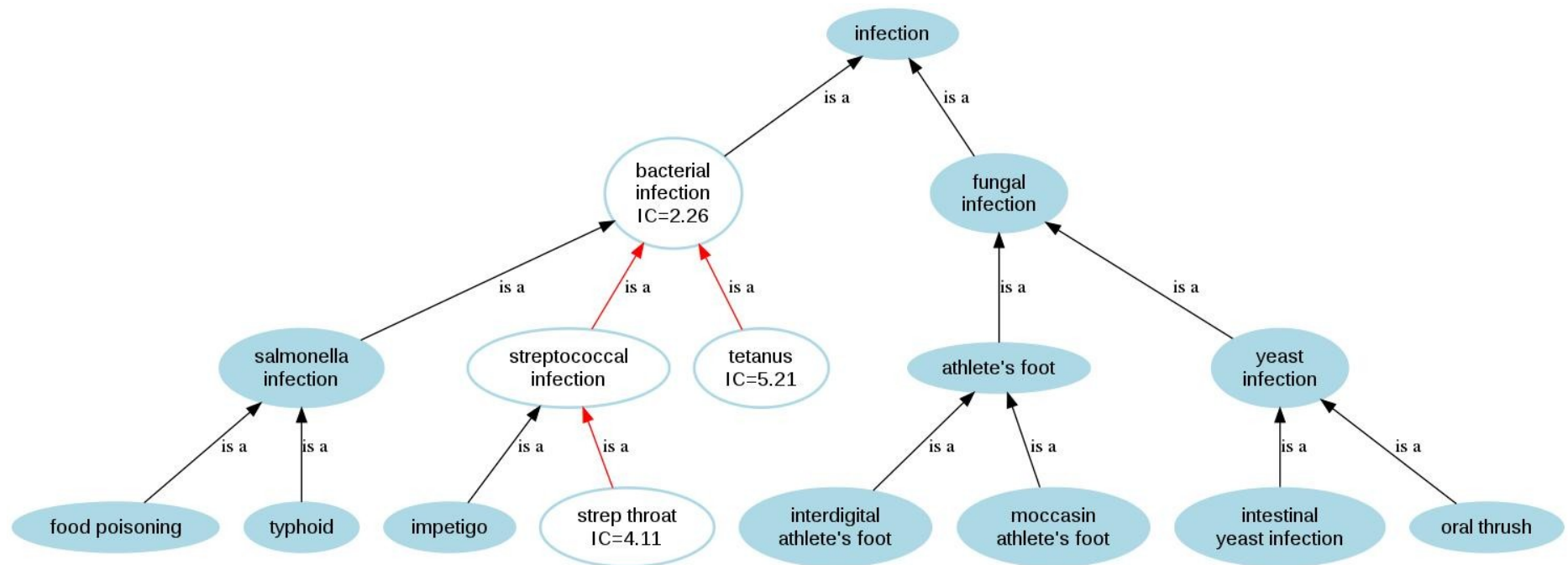
- $lin(a,b) = \frac{\text{-----}}{IC (a) + IC (b)}$

- Look familiar?

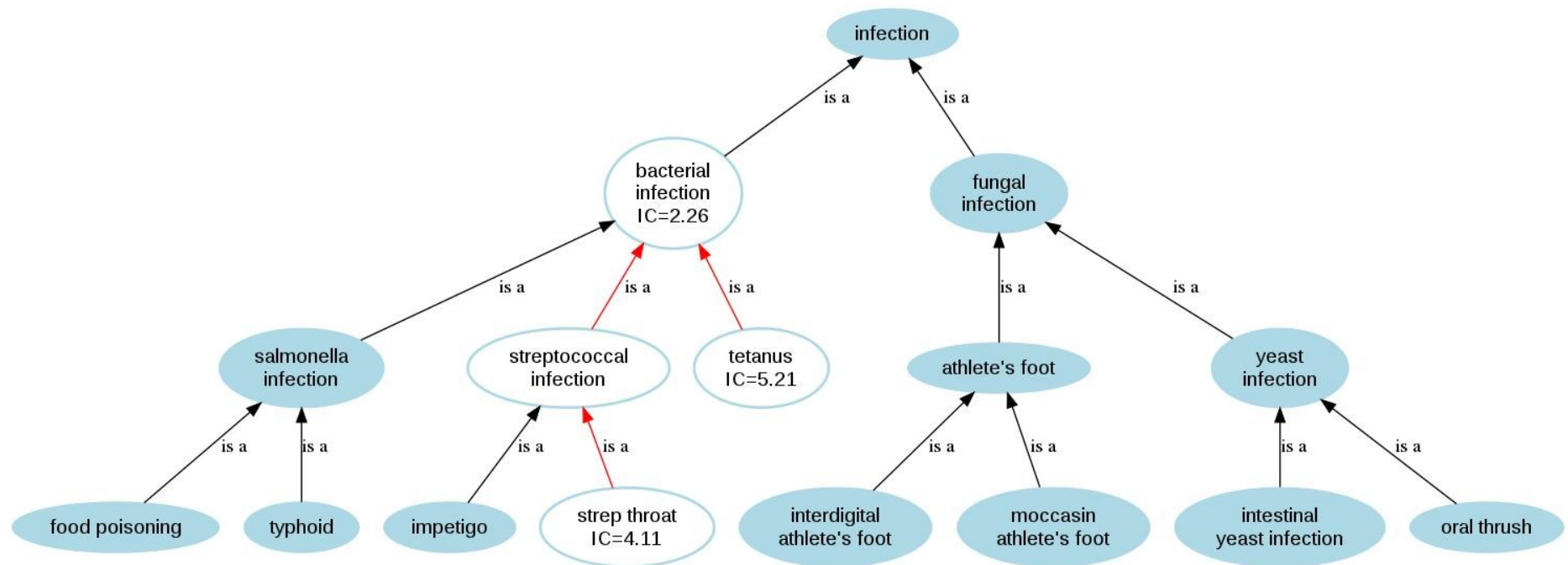
$$2* \text{depth} (LCS (a,b))$$

- $wup(a,b) = \frac{\text{-----}}{\text{depth}(a) + \text{depth} (b)}$

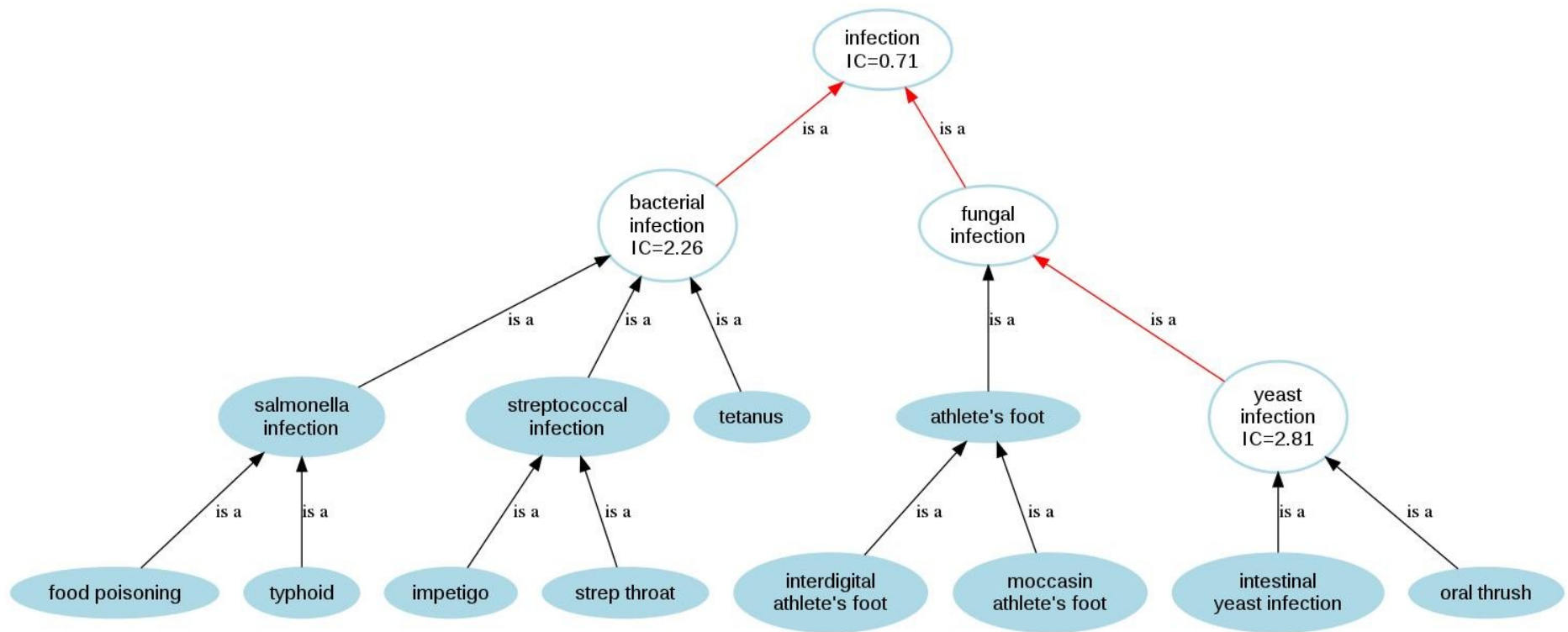
$$\text{lin}(\text{strep throat}, \text{tetanus}) = 2 * 2.26 / (5.21 + 4.11) = 0.485$$



$$\text{lin}(\text{strep throat}, \text{tetanus}) = 2 * 2.26 / (5.21 + 4.11) = 0.485$$



$$\text{lin}(\text{bacterial infection}, \text{yeast infection}) = 2 * 0.71 / (2.26 + 2.81) = 0.280$$



?

- Lin says that *strep throat* and *tetanus* (.49) are more similar than are *bacterial infection* and *yeast infection* (.28)
- Wu and Palmer say that *strep throat* and *tetanus* (.57) are more similar than are *bacterial infection* and *yeast infection* (.4)
- Path says that *strep throat* and *tetanus* (.25) are equally similar as are *bacterial infection* and *yeast infection* (.25)

What about concepts not connected via is-a relations?

- Connected via other relations?
 - Part-of, treatment-of, causes, etc.
- Not connected at all?
 - In different sections (axes) of an ontology (infections and treatments)
 - In different ontologies entirely (SNOMEDCT and FMA)
- Relatedness!
 - Use definition information
 - No is-a relations so can't be similarity

Measures of relatedness

- Path based
 - Hirst & St-Onge, 1998 (hso)
- Definition based
 - Lesk, 1986
 - Adapted lesk (lesk)
 - Banerjee & Pedersen, 2003
- Definition + corpus
 - Gloss Vector (vector)
 - Patwardhan & Pedersen, 2006

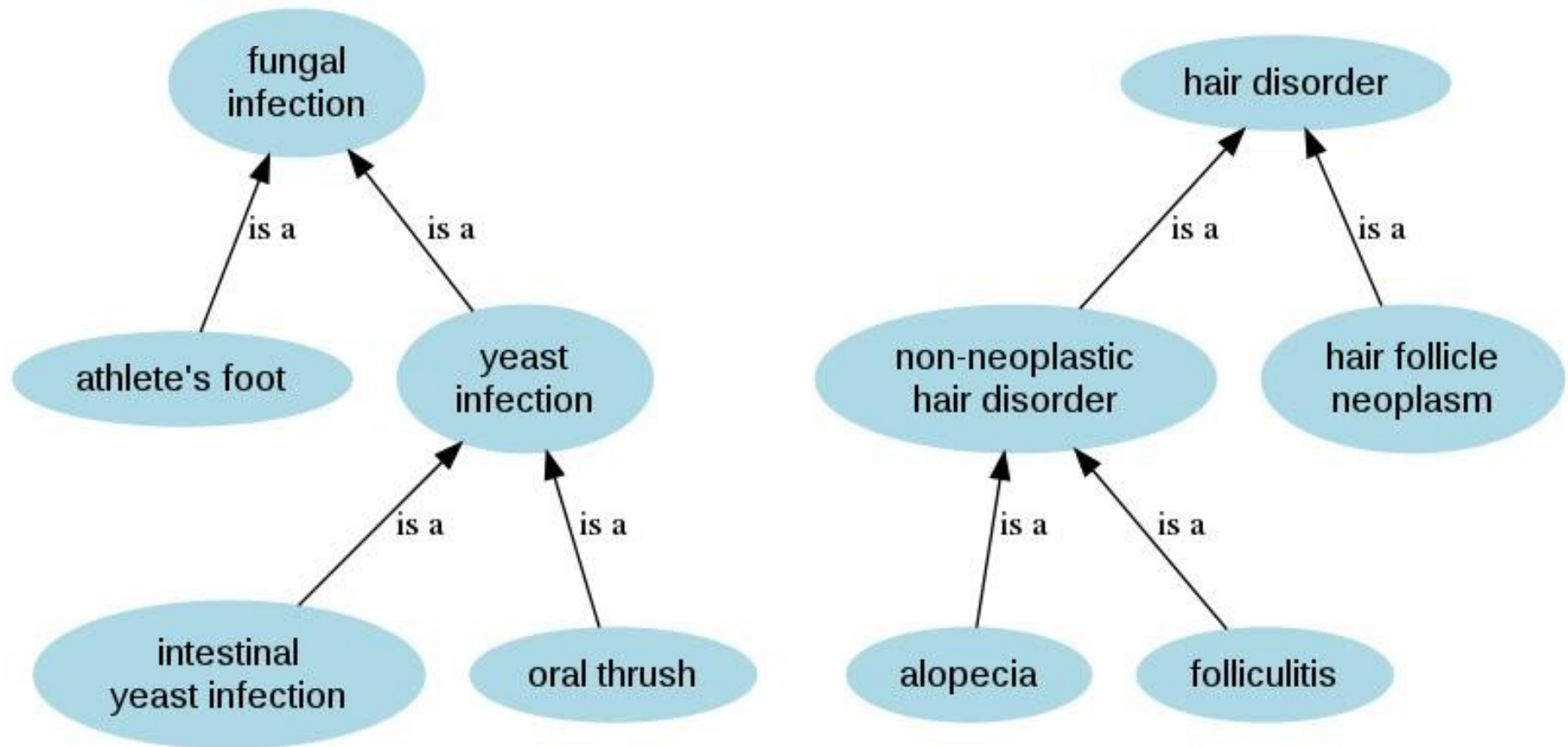
Path based relatedness

- Ontologies include relations other than is-a
- These can be used to find shortest paths between concepts
 - However, a path made up of different kinds of relations can lead to big semantic jumps
 - Aspirin treats headaches which are a symptom of the flu which can be prevented by a flu vaccine which is recommend for children
 - so aspirin and children are related ??

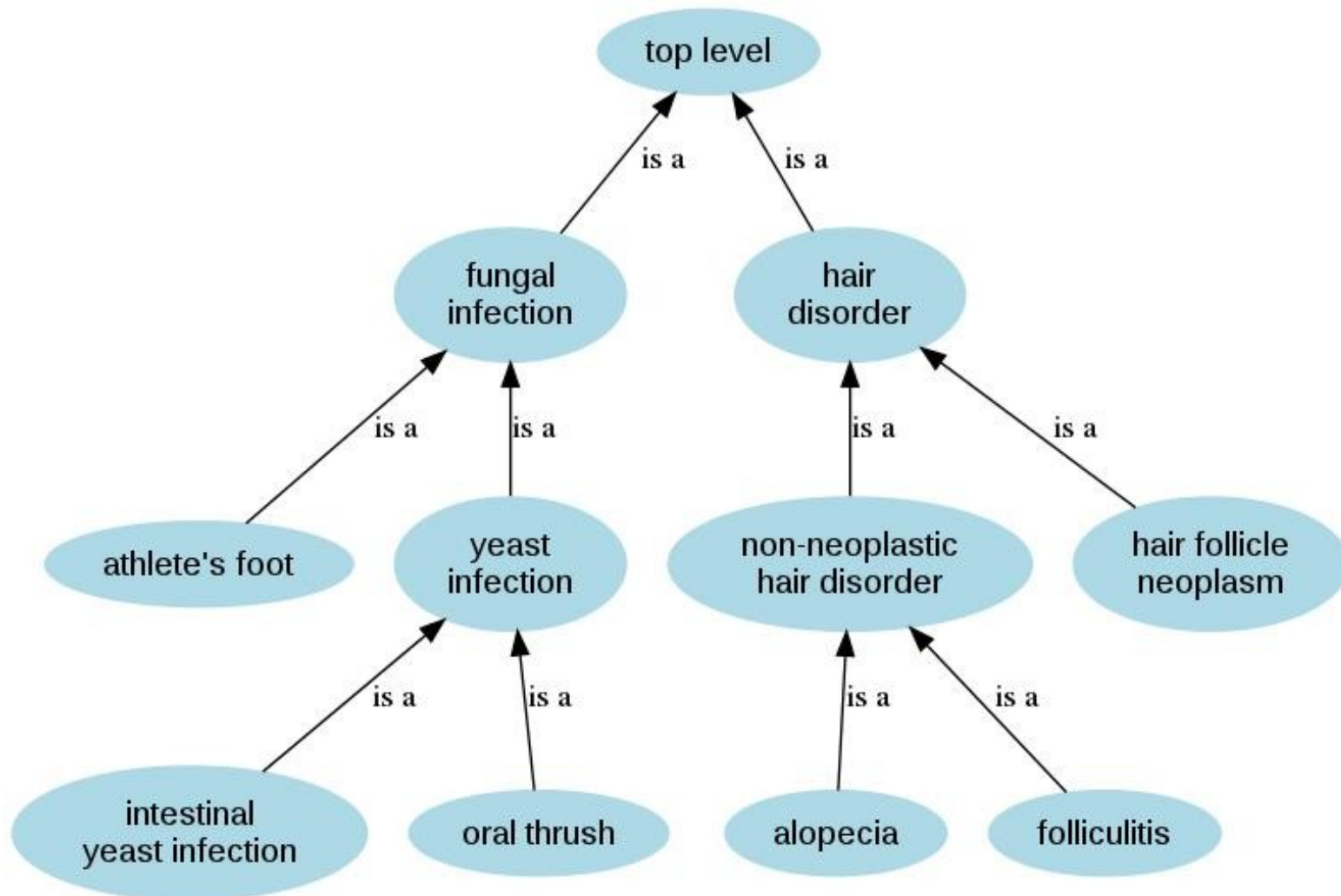
Measuring relatedness with definitions

- Related concepts defined using many of the same terms
- But, definitions are short, inconsistent
- Concepts don't need to be connected via relations or paths to measure them
 - Lesk, 1986
 - Adapted Lesk, Banerjee & Pedersen, 2003

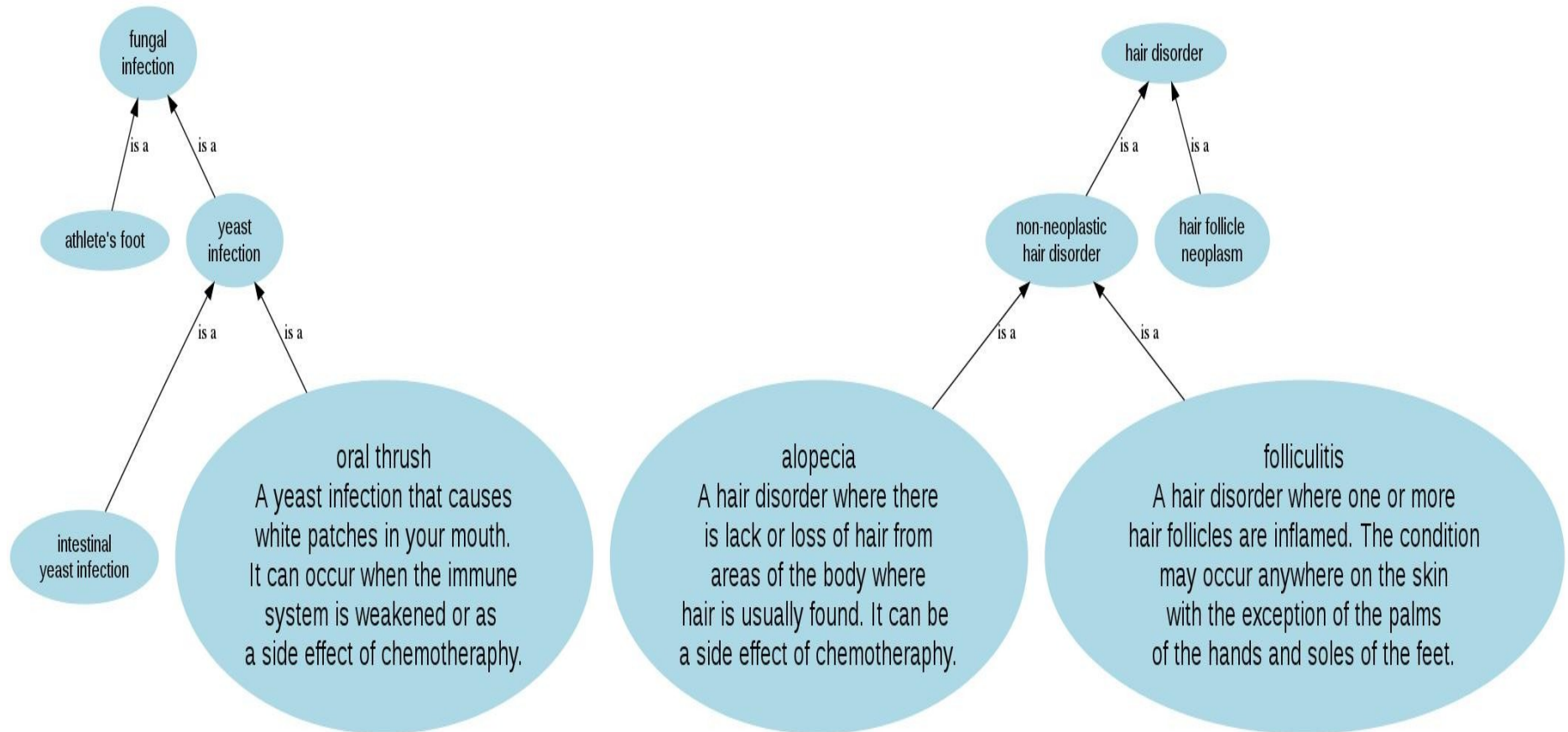
Two separate ontologies...



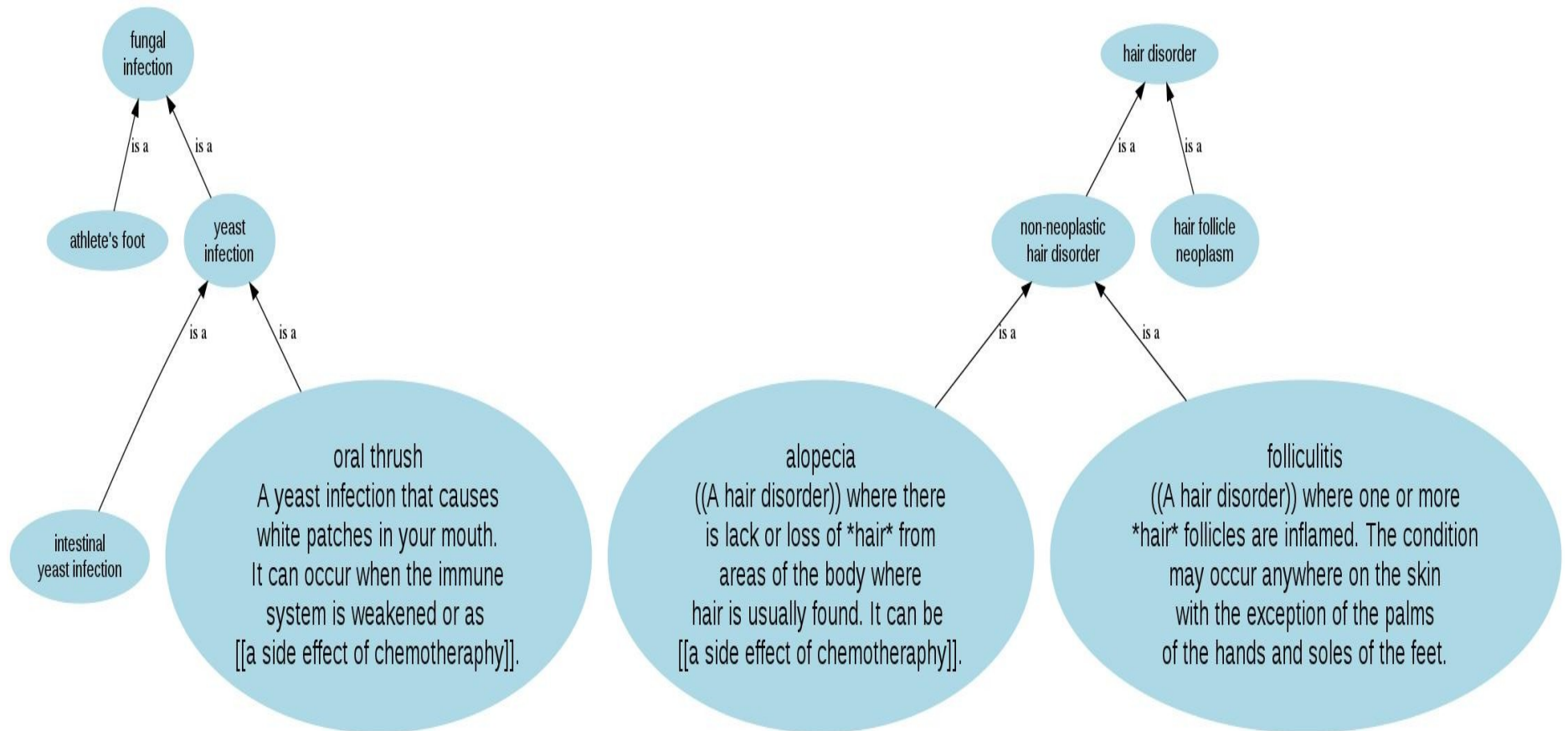
Could join them together ... ?



Each concept has definition



Find overlaps in definitions...



Overlaps

- Oral Thrush and Alopecia
 - side effect of chemotherapy
 - Can't see this in structure of is-a hierarchies
 - Oral thrush and folliculitis just as similar
- Alopecia and Folliculitis
 - hair disorder & hair
 - Reflects structure of is-a hierarchies
 - If you start with text like this maybe you can build is-a hierarchies automatically!
 - Another tutorial...

Lesk and Adapted Lesk

- Lesk, 1986 : measure overlaps in definitions to assign senses to words
 - The more overlaps between two senses (concepts), the more related
- Banerjee & Pedersen, 2003, Adapted Lesk
 - Augment definition of each concept with definitions of related concepts
 - Build a super gloss
 - Increase chance of finding overlaps
- Implemented in UMLS::Similarity as lesk

The problem with definitions ...

- Definitions contain variations of terminology that make it impossible to find exact overlaps
- Alopecia : ... a **result** of cancer treatment
- Thrush : ... a **side effect** of chemotherapy
 - Real life example, I modified the alopecia definition to work better with Lesk!!!
 - NO MATCHES!!
- How can we see that “result” and “side effect” are similar, as are “cancer treatment” and “chemotherapy” ?

Gloss Vector Measure of Semantic Relatedness

- Rely on co-occurrences of terms
 - Terms that occur within some given number of terms of each other
- Allows for a fuzzier notion of matching
- Exploits second order co-occurrences
 - Friend of a friend relation
 - Suppose *cancer_treatment* and *chemotherapy* don't occur in text with each other. But, suppose that “*survival*” occurs with each.
 - *cancer_treatment* and *chemotherapy* are second order co-occurrences via “*survival*”

Gloss Vector Measure of Semantic Relatedness

- Replace words or terms in definitions with vector of co-occurrences observed in corpus
- Defined concept now represented by an averaged vector of co-occurrences
- Measure relatedness of concepts via cosine between their respective vectors
- Patwardhan and Pedersen, 2006
 - Inspired by Schutze, 1998
- Implemented in UMLS::Similarity as vector

Thank you!

- <http://umls-similarity.sourceforge.net>
 - Tutorial slides
 - Links to web interfaces
 - Software downloads
 - Mailing list (join!!)
- Next – Using UMLS::Similarity software !

References

- S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, pages 805-810, Acapulco, August 2003.
- J. Caviedes and J. Cimino. Towards the development of a conceptual distance metric for the UMLS. Journal of Biomedical Informatics, 37(2):77-85, April 2004.
- J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings on International Conference on Research in Computational Linguistics, pages 19-33, Taiwan, 1997.
- C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, WordNet: An electronic lexical database, pages 265-283. MIT Press, 1998.

References

- M.E. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation, pages 24-26. ACM Press, 1986.
- D. Lin. An information-theoretic definition of similarity. In Proceedings of the International Conference on Machine Learning, Madison, August 1998.
- H.A. Nguyen and H. Al-Mubaid. New ontology-based semantic similarity measure for the biomedical domain. In Proceedings of the IEEE International Conference on Granular Computing, pages 623-628, Atlanta, GA, May 2006.
- S. Patwardhan and T. Pedersen. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In Proceedings of the EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together, pages 1-8, Trento, Italy, April 2006.

References

- R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1):17-30, 1989.
- P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448-453, Montreal, August 1995.
- H. Schütze. Automatic word sense discrimination. Computational Linguistics, 24(1):97-123, 1998.
- J. Zhong, H. Zhu, J. Li, and Y. Yu. Conceptual graph matching for semantic search. Proceedings of the 10th International Conference on Conceptual Structures, pages 92-106, 2002

Supplemental Materials

- Semantic Similarity for the Gene Ontology
 - Various measures for GO :
 - http://www.geneontology.org/GO.tools_by_type.semantic_similarity.shtml

Supplemental Materials

- WordNet::Similarity
 - Predecessor of UMLS::Similarity
 - Developed at University of Minnesota, Duluth 2001-2006
 - Based on English lexical database WordNet
 - <http://wordnet.princeton.edu>
 - <http://wn-similarity.sourceforge.net>

Supplemental Materials

- WebServices::UMLSKS::Similarity
 - HSO implementation for UMLS
 - Developed at University of Minnesota, Duluth 2010-2012, ongoing
 - <http://search.cpan.org/dist/WebService-UMLSKS-Similarity/>

Introducing UMLS::Interface and UMLS::Similarity (without tears)

Bridget T. McInnes

bthomson@umn.edu
<http://www.tc.umn.edu/~bthomson>

Just in case

Kleenex is still available in the back of the room

Outline

- Unified Medical Language System
- UMLS::Interface
 - Backbone of UMLS::Similarity
- UMLS::Similarity

Unified Medical Language System

- UMLS
 - Metathesaurus
 - Semantic Network
 - SPECIALIST LEXICON

Metathesaurus

- ~1.7 million biomedical and clinical concepts; integrated semi-automatically
 - CUIs (Concept Unique Identifiers)
 - Hierarchical Relations
 - PAR/CHD (parent/child)
 - RB/RN (broader/narrower)
 - Non-hierarchical Relations
 - SIB (sibling)
 - RO (other relation)
 - Definitional information

Metathesaurus Sources

- Foundational Model of Anatomy (FMA)
- Medical Subject Headings (MSH)
- SNOMED Clinical Terms (SNOMEDCT)

UMLS::Interface

UMLS::Interface

- Perl interface to the UMLS present locally in a MySQL database.
- Backbone to UMLS::Similarity
- Main purpose is to return information about CUIs
 - Path information
 - Definitional information

Using UMLS::Interface

- Two ways to interact with UMLS::Interface
 - API
 - command line programs

Using UMLS::Interface

- Two ways to interact with UMLS::Interface
 - API
 - command line interface programs

Nice API examples in the
UMLS::Similarity package

Using UMLS::Interface

- Two ways to interact with UMLS::Interface
 - API
 - command line interface programs

Using UMLS::Interface

- Two ways to interact with UMLS::Interface
 - API
 - command line interface programs: 26
 - findPathToRoot.pl
 - findShortestPath.pl
 - getCuiDef.pl
 - findLeastCommonSubsumer.pl
 - getChildren.pl
 - getParents.pl
 - getRelated.pl
 - Ect

Using UMLS::Interface

- Two ways to interact with UMLS::Interface
 - API
 - command line interface programs: 26
 - findPathToRoot.pl
 - findShortestPath.pl
 - getCuiDef.pl
 - findLeastCommonSubsumer.pl
 - getChildren.pl
 - getParents.pl
 - getRelated.pl
 - Ect

findPathToRoot.pl

findPathToRoot.pl <CUI | Term>

findPathToRoot.pl

findPathToRoot.pl “bacterial infection”

findPathToRoot.pl

findPathToRoot.pl "bacterial infection"

The paths between bacterial infection (C0004623) and the root: =>

C00000000 (**UMLS ROOT**)

C1135584 (msh)

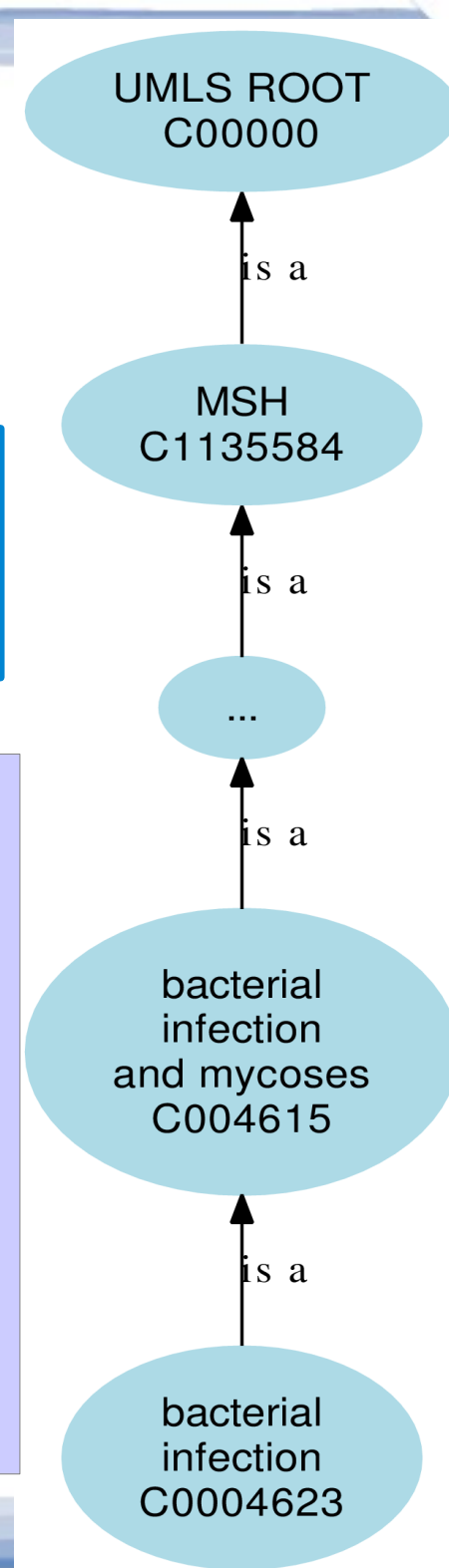
C1256739 (mesh descriptors)

C1256741 (index medicus descriptor)

C0012674 (diseases)

C0004623 (bacterial infections and mycoses)

C0004623 (bacterial infextion)



findPathToRoot.pl

findPathToRoot.pl "bacterial infection"

The paths between bacterial infection (C0004623) and the root: =>

C00000000 (**UMLS ROOT**)

C1135584 (msh)

C1256739 (mesh descriptors)

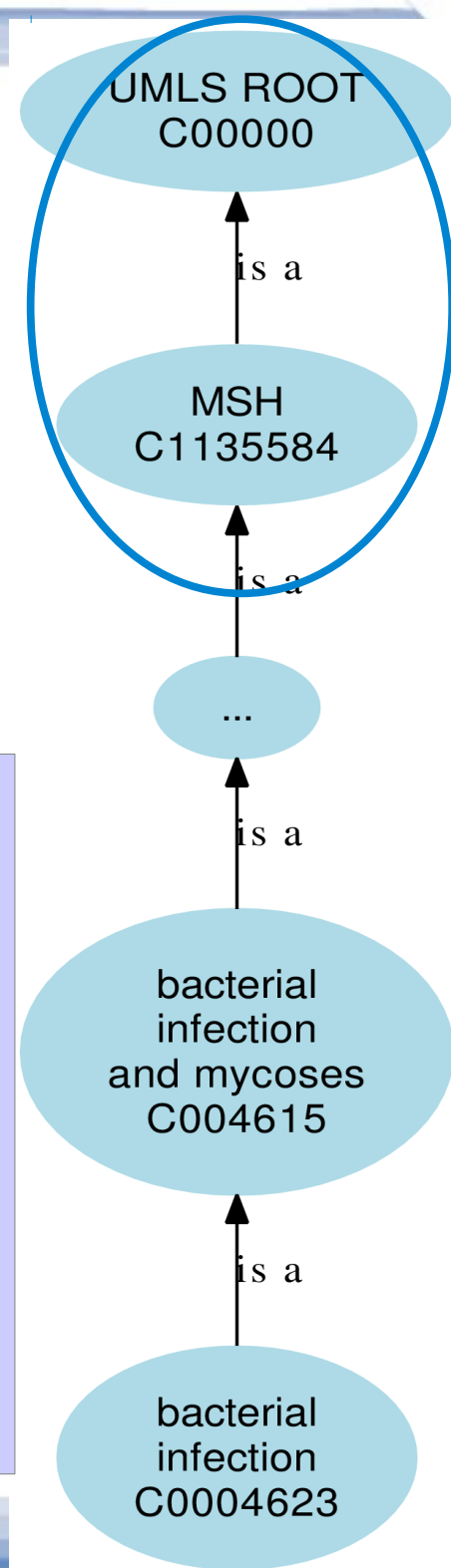
C1256741 (index medicus descriptor)

C0012674 (diseases)

C0004623 (bacterial infections and mycoses)

C0004623 (bacterial infexction)

DEFAULT:
MSH
PAR/CHD



--config option

CONFIG FILE NAMED 'config'

```
SAB :: include <source1>,<source2>  
REL :: include <relation1>,<relation2>
```

findPathToRoot.pl with --config

```
findPathToRoot.pl  "bacterial infection"  
                  --config config
```

CONFIG FILE NAMED 'config'

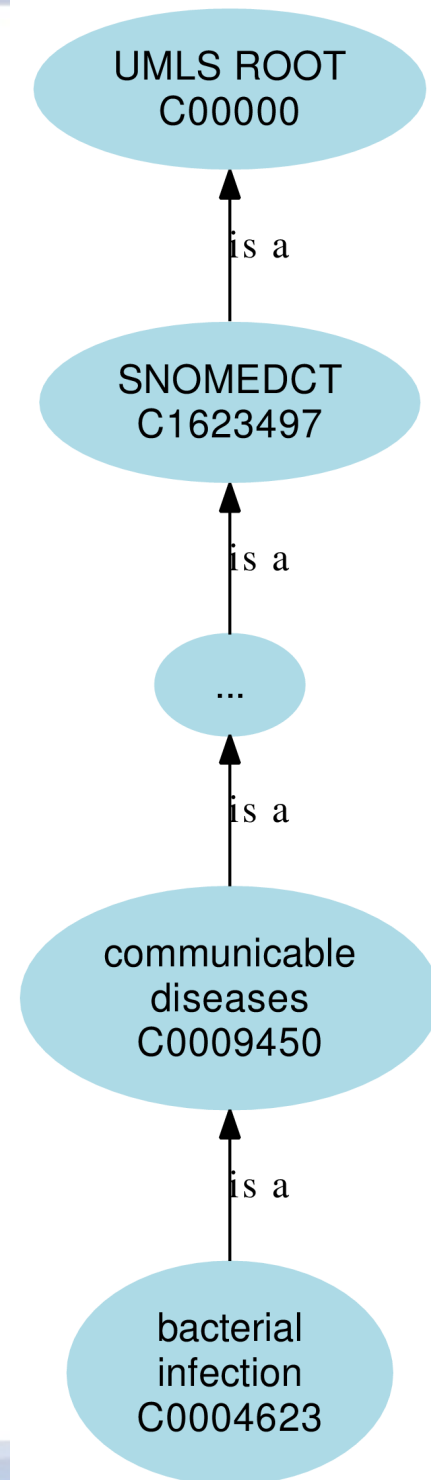
```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

findPathToRoot.pl with --config

```
findPathToRoot.pl  "bacterial infection"  
                  --config config
```

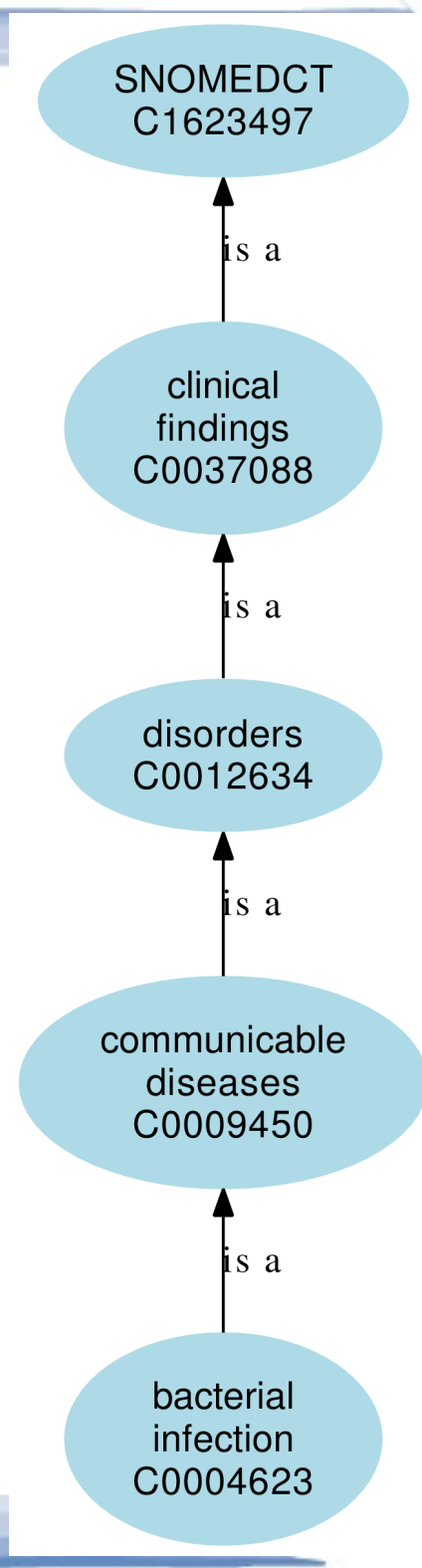
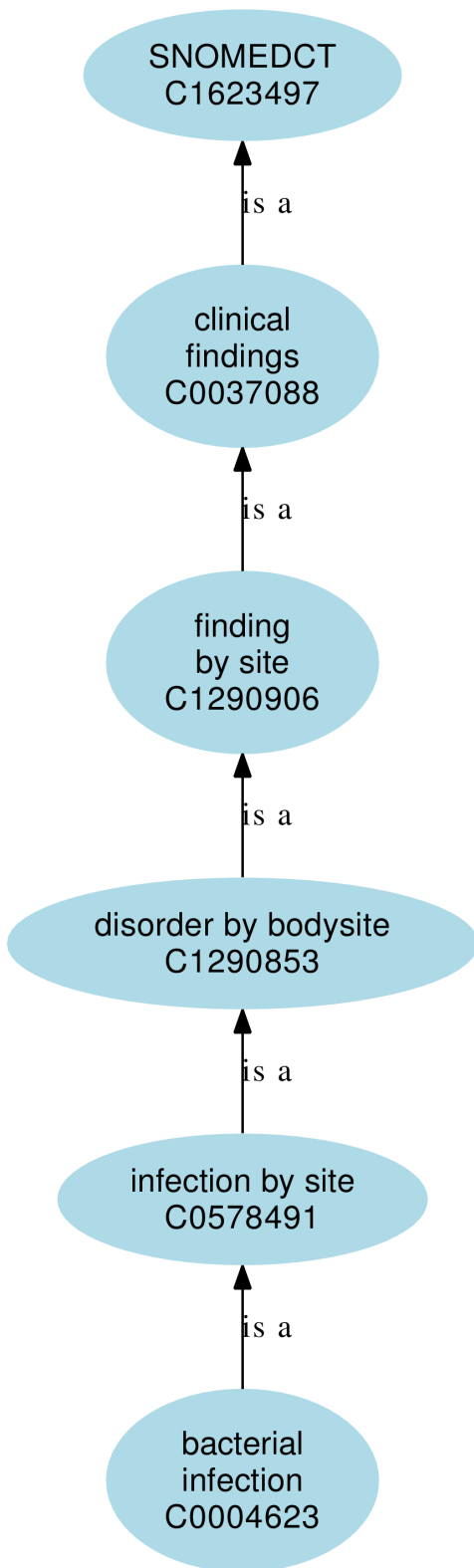
The paths between bacterial infection (C0004623)
and the root:

C0000000 (**UMLS ROOT**)
C1623497 (SNOMEDCT)
C0037088 (clinical findings)
C0012634 (disorders)
C0009450 (communicable diseases)
C0004623 (bacterial infection)



findPathToRoot.pl multiple paths

ALL POSSIBLE PATHS
ARE RETURNED



findShortestPath.pl

```
findShortestPath.pl <CUI or Term>  
                  < CUI or Term>
```


findShortestPath.pl

```
findShortestPath.pl <CUI or Term>  
                  < CUI or Term>
```

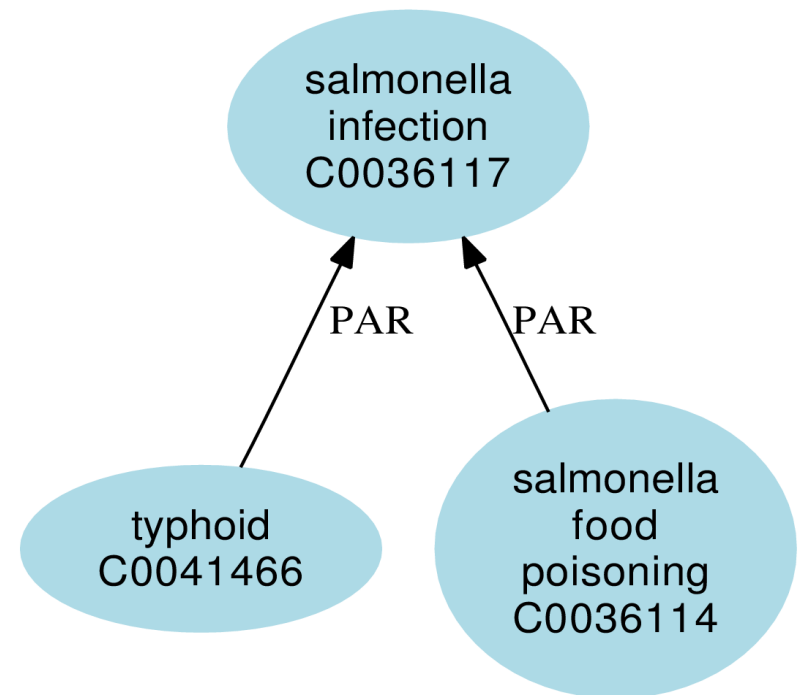
**DEFAULT: MSH using the
PAR/CHD relations**

findShortestPath.pl

```
findShortestPath.pl "salmonella food poisoning"  
"typhoid"  
-config config
```

CONFIG FILE NAMED 'config'

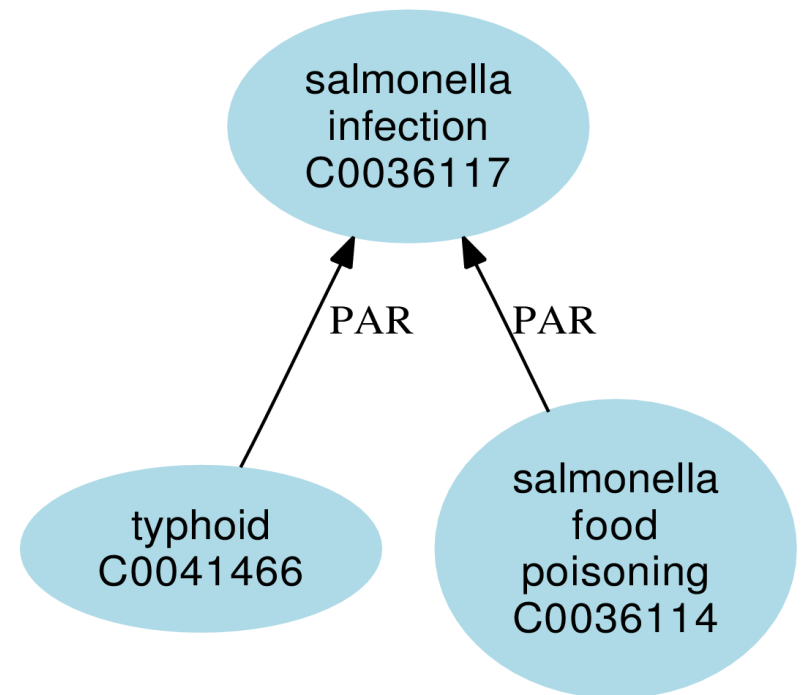
```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```



findShortestPath.pl

findShortestPath.pl “salmonella food poisoning”
“typhoid”
–config config

The shortest path between salmonella food poisoning (C0036114) and typhoid (C0041466):
=> C0036114 (salmonella food poisoning)
=> C0036117 (salmonella infection)
=> C0041466 (fever, typhoid)

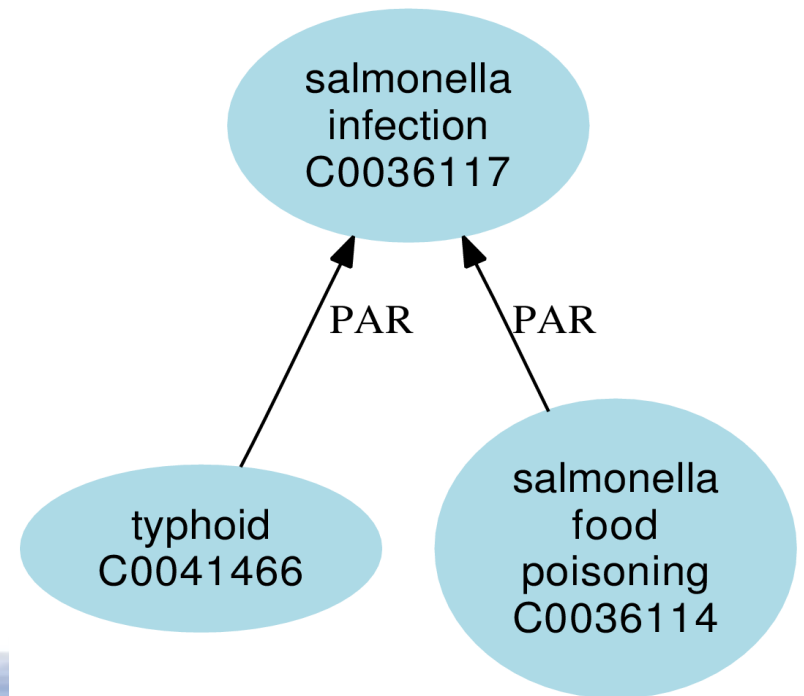


findShortestPath.pl

findShortestPath.pl “salmonella food poisoning”
“typhoid”
–config config
– info

The shortest path between salmonella food poisoning (C0036114) and the typhoid (C0041466):

=> C0036114 (salmonella food poisoning)
=> PAR (SNOMEDCT)
=> C0036117 (salmonella infection)
=> CHD (SNOMEDCT)
=> C0041466 (fever, typhoid)

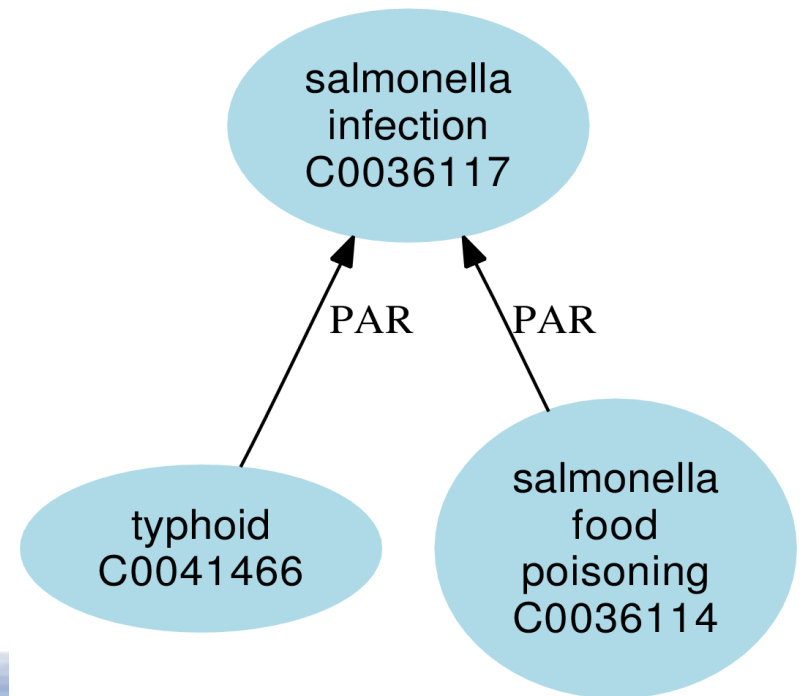


findLeastCommonSubsumer.pl

```
findLeastCommonSubsumer.pl  
    "salmonella food poisoning"  
    "typhoid"  
    -config config
```

CONFIG FILE NAMED 'config'

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```



findLeastCommonSubsumer.pl

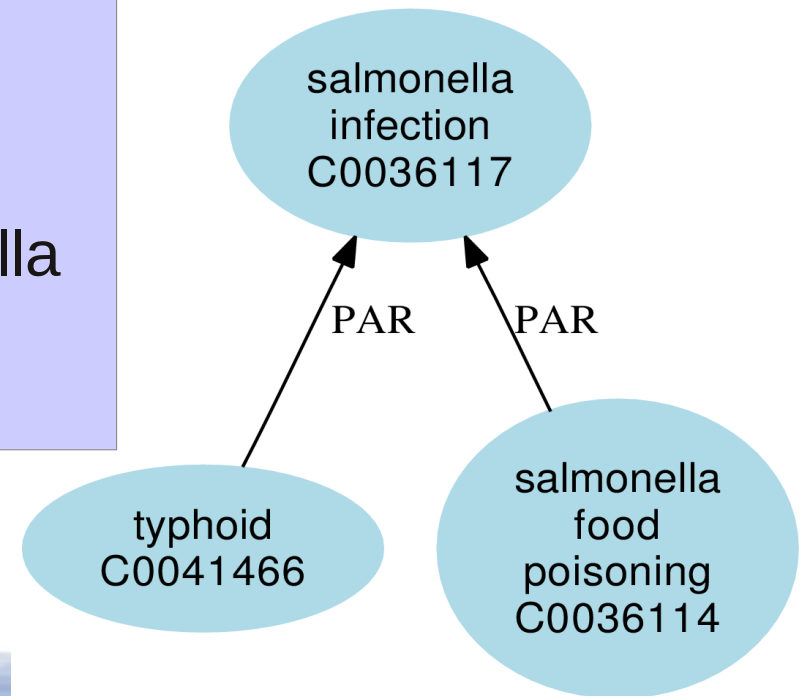
findLeastCommonSubsumer.pl

“salmonella food poisoning”

“typhoid”

–config config

The least common subsumer between salmonella food poisoning (C0036114) and the typhoid (C0041466) is salmonella infection (C0036117)



getCuiDef.pl

getCuiDef.pl <CUI or Term>

getCuiDef.pl

getCuiDef.pl typhoid

The definition(s) of typhoid (C0041466):

1. acute systemic febrile infection caused by *Salmonella typhi*.
2. an acute systemic febrile infection caused by *SALMONELLA TYPHI*, a serotype of *SALMONELLA ENTERICA*.

getCuiDef.pl with --sab

```
getCuiDef.pl typhoid --sab
```

The definition(s) of typhoid (C0041466):

1. CSP acute systemic febrile infection caused by *Salmonella typhi*.
2. MSH an acute systemic febrile infection caused by *SALMONELLA TYPHI*, a serotype of *SALMONELLA ENTERICA*.

UMLS::Similarity

UMLS::Similarity

- A suite of perl modules: implement a number of semantic similarity and relatedness measures
- Similarity measures:
 - Use UMLS path information obtained by UMLS:Interface
- Relatedness measures:
 - The UMLS definition information obtained by UMLS::Interface

Measures

- Path-based similarity
 - Use only the path information between the two concepts
- Information content-based similarity
 - Incorporate the probability of the concept occurring in some text
- Relatedness measures
 - Use contextual information about a concept

Path-based Similarity Measures

- Conceptual Distance measure
 - Rada, et al (1989)
 - Caviedes and Cimino (2004)
- Wu and Palmer (1994)
- Leacock and Chodorow (1998)
- Zhong, et al (2002)
- Ngyuen and Al-Mubaid (2006)

Information-content based Similarity Measures

- Resnik (1995)
- Lin (1997)
- Jiang and Conrath (1997)

Relatedness Measures

- Adapted Lesk
 - Banerjee and Pedersen (2003)
- Gloss Vector
 - Patwardhan and Pedersen (2006)

Using UMLS::Similarity

- Three ways to interact with UMLS::Similarity
 - API
 - command line
 - web interface
 - <http://atlas.ahc.umn.edu/>
 - <http://maraca.d.umn.edu/>

Using UMLS::Similarity

- Three ways to interact with UMLS::Similarity
 - API
 - command line
 - web interface
 - <http://atlas.ahc.umn.edu/>
 - <http://maraca.d.umn.edu/>

umls-similarity.pl

- Main program in UMLS::Similarity
- At its most basic:
 - Takes two terms or CUIs as input
 - Returns similarity between them

```
umls-similarity.pl <CUI or Term> <CUI or Term>
```

umls-similarity.pl using defaults

```
umls-similarity.pl tetanus salmonella
```

umls-similarity.pl using defaults

```
umls-similarity.pl tetanus salmonella
```

```
0.0769<>tetanus(C0039614)<>salmonella(C0036111)
```

umls-similarity.pl using defaults

```
umls-similarity.pl tetanus salmonella
```

```
0.0769<>tetanus(C0039614)<>salmonella(C0036111)
```

Source: MSH

Relations: PAR/CHD

Measure: path = 1 / length of the shortest
path between the two
concepts

umls-similarity.pl with --config

```
umls-similarity.pl tetanus salmonella  
--config config
```

umls-similarity.pl with --config

```
umls-similarity.pl tetanus salmonella  
--config config
```

CONFIG FILE NAMED 'config'

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

umls-similarity.pl with --config

```
umls-similarity.pl tetanus salmonella  
--config config
```

```
0.0714<>tetanus(C0039614)<>salmonella(C0036111)
```

CONFIG FILE NAMED 'config'

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

umls-similarity.pl with --config

```
umls-similarity.pl tetanus salmonella  
--config config
```

0.0714<>tetanus(C0039614)<>salmonella(C0036111)

MSH



SNOMEDCT



0.0769<>tetanus(C0039614)<>salmonella(C0036111)₄₇

Viewing sources in combination

```
umls-similarity.pl tetanus salmonella  
-config config
```

```
0.0714<>tetanus(C0039614)<>salmonella(C0036111)
```

CONFIG FILE NAMED 'config'

```
SAB :: include SNOMEDCT, MSH  
REL :: include PAR, CHD
```

Adding RB/RN relations

```
umls-similarity.pl tetanus salmonella  
-config config
```

```
0.0714<>tetanus(C0039614)<>salmonella(C0036111)
```

CONFIG FILE NAMED 'config'

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD, RB, RN
```

Be careful

- Two reasons:
 - Sources were created for different purposes therefore have different granularity in their path information
 - UMLS::Similarity builds an index of the path to root information of all the CUIs in the source and relations specified in the config file
 - This can get large!

Using the –realtime option

```
umls-similarity.pl tetanus salmonella  
                  –config config  
                  – realtime
```

```
0.0714<>tetanus(C0039614)<>salmonella(C0036111)
```

CONFIG FILE NAMED 'config'

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD, RB, RN
```


For the brave – using the entire UMLS

```
umls-similarity.pl tetanus salmonella  
                  –config config  
                  – realtime
```

```
0.0714<>tetanus(C0039614)<>salmonella(C0036111)
```

CONFIG FILE NAMED 'config'

```
SAB :: include UMLS_ALL  
REL :: include PAR, CHD
```

umls-similarity.pl with –infile option

```
umls-similarity.pl --infile <input file>
```

input file

```
C0035078<>C0035078  
C0018787<>C0027061  
C0026269<>C0004238  
tetanus<>salmonella  
typhoid<>salmonella  
strep throat<>impetigo  
C0004623<>tetanus  
typhoid<>C0018787  
C0026269<>impetigo
```

umls-similarity.pl with -infile option

```
umls-similarity.pl --infile <input file>
```

output

```
1.0000<>C0035078(Failure, Kidney)<>C0035078(Failure, Kidney)
0.5000<>C0018787(Heart, NOS)<>C0027061(Myocardium, NOS)
0.2000<>C0026269(Stenosis)<>C0004238(Fibrillation, Atrial)
0.0769<>tetanus(C0039614)<>salmonella(C0036111)
0.0714<>typhoid(C0041466)<>salmonella(C0036111)
-1.0000<>strep throat<>impetigo
0.2500<>Infections, Bacterial(C0004623)<>tetanus(C0039614)
0.0909<>typhoid(C0041466)<>Heart, NOS(C0018787)
0.1000<>Stenosis(C0026269)<>impetigo(C0021099)
```

umls-similarity.pl with –infile option

```
umls-similarity.pl --infile <input file>
```

output

```
1.0000<>C0035078(Failure, Kidney)<>C0035078(Failure, Kidney)
0.5000<>C0018787(Heart, NOS)<>C0027061(Myocardium, NOS)
0.2000<>C0026269(Stenosis)<>C0004238(Fibrillation, Atrial)
0.0769<>tetanus(C0039614)<>salmonella(C0036111)
0.0714<>typhoid(C0041466)<>salmonella(C0036111)
-1.0000<>strep throat<>impetigo
0.2500<>Infections, Bacterial(C0004623)<>tetanus(C0039614)
0.0909<>typhoid(C0041466)<>Heart, NOS(C0018787)
0.1000<>Stenosis(C0026269)<>impetigo(C0021099)
```

--measure options for path-based measures

- path: simple path
- cdist: conceptual distance
- wup: Wu and Palmer (1994)
- lch: Leacock and Chodorow (1998)
- zhong: Zhong, et al (2002)
- nam: Ngyuen and Al-Mubaid (2006)

--measure options for path-based measures

- path: simple path
- cdist: conceptual distance
- wup: Wu and Palmer (1994)
- lch: Leacock and Chodorow (1998)
- zhong: Zhong, et al (2002)
- nam: Ngyuen and Al-Mubaid (2006)

```
umls-similarity.pl <CUI or Term> <CU or Term>  
                  --measure <measures>
```

umls-similarity.pl with --measure

```
umls-similarity.pl tetanus salmonella  
--measure wup
```

```
0.3636<>tetanus(C0039614)<>salmonella(C0036111)
```

umls-similarity.pl with --measure

```
umls-similarity.pl tetanus salmonella  
--measure wup
```

```
0.3636<>tetanus(C0039614)<>salmonella(C0036111)
```

```
umls-similarity.pl tetanus salmonella  
--measure lch
```

```
1.1239<>tetanus(C0039614)<>salmonella(C0036111)
```


--measure options for IC-based measures

- res: Resnik (1996)
- lin: Lin (1997)
- jcn: Jiang and Conrath (1997)

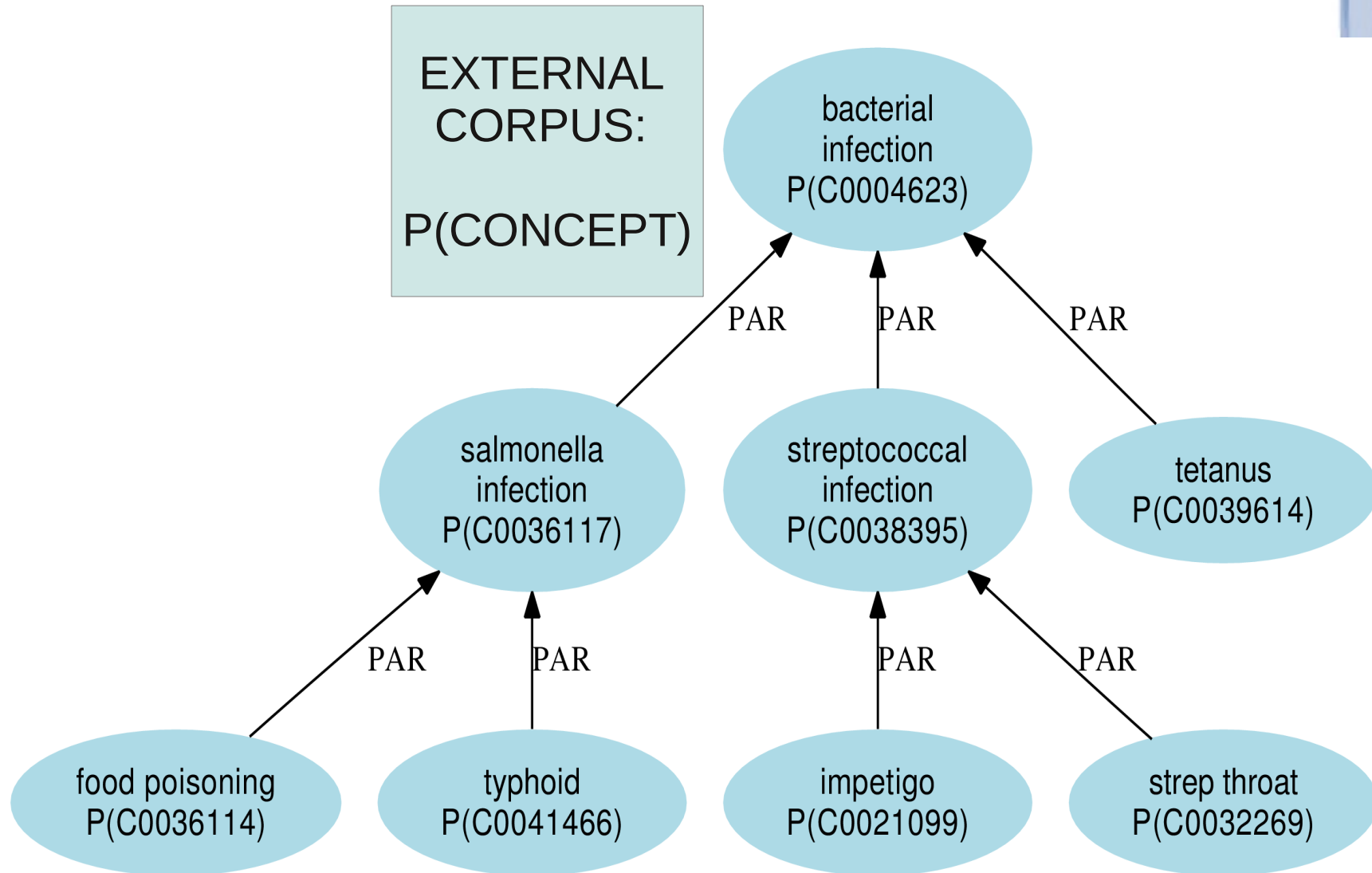
```
umls-similarity.pl tetanus salmonella  
-measure lin
```

```
0.3636<>tetanus(C0039614)<>salmonella(C0036111)
```

Information Content

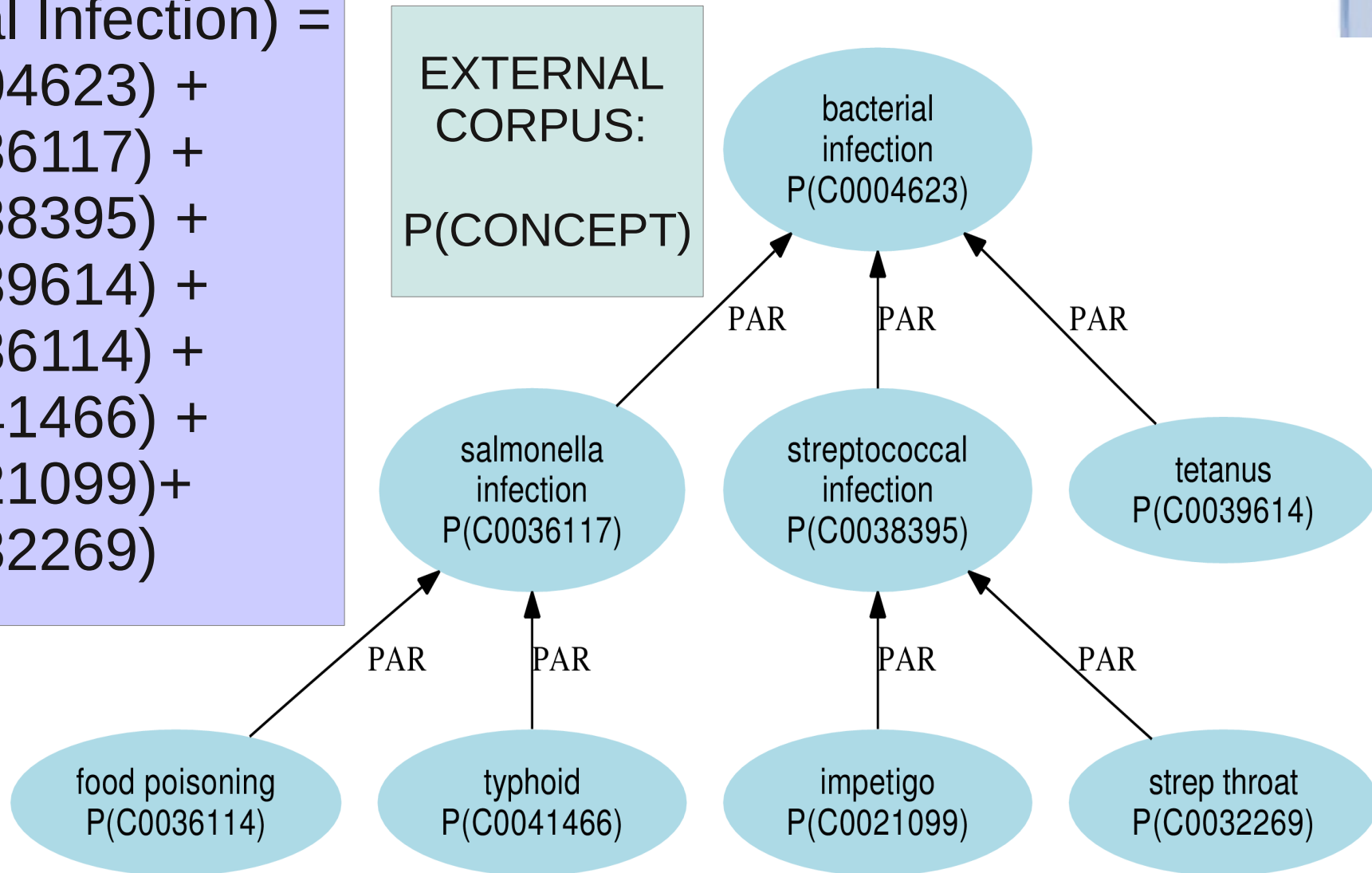
- $IC = -\log(P(\text{concept}))$
- $P(\text{concept})$:
 - Calculated by summing the probability of the concept seen a corpus with the probability of the concept's descendants
 - Probabilities obtained from external corpus

Example: probability of hand



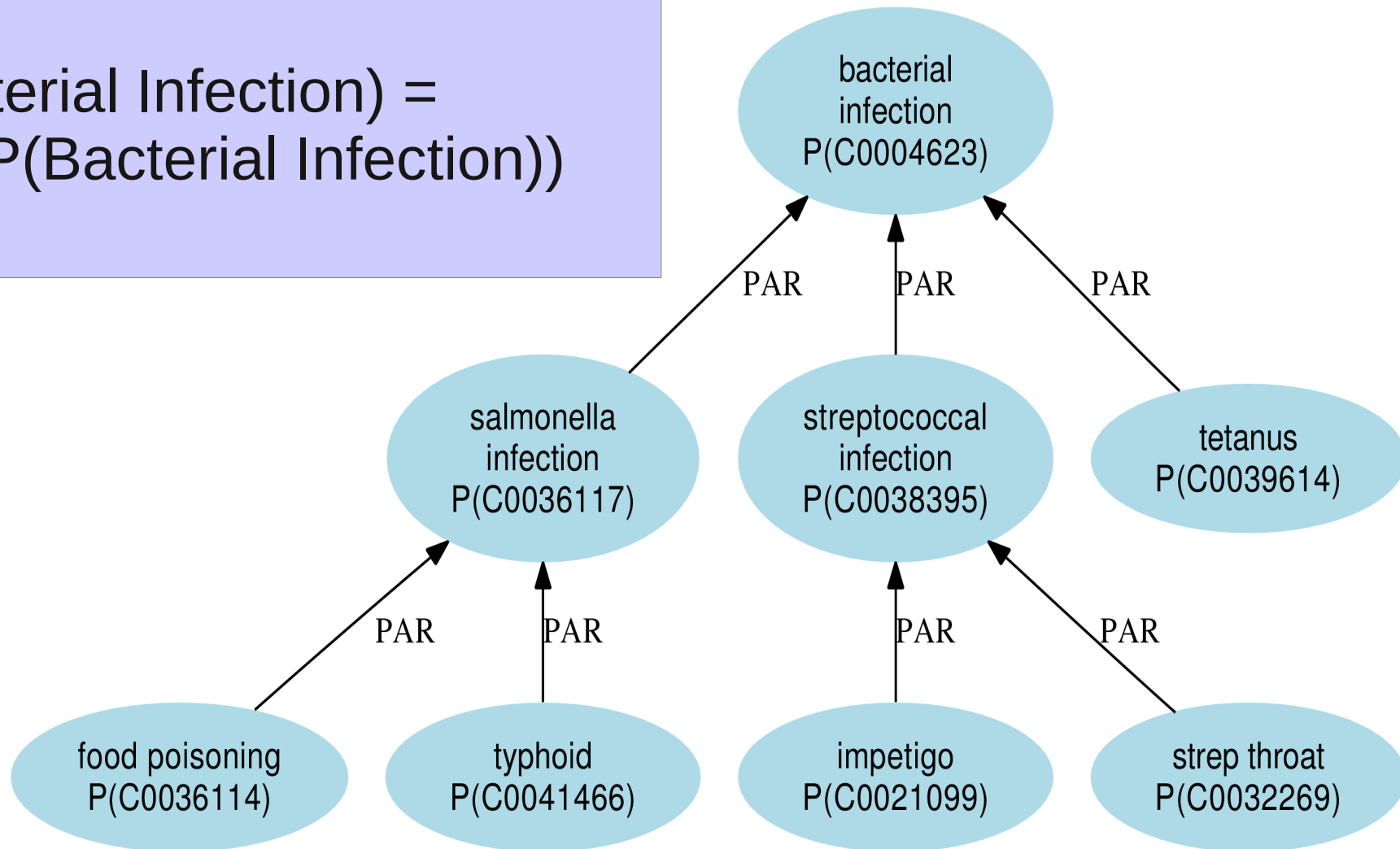
Example: probability of hand

$$\begin{aligned} P(\text{Bacterial Infection}) = & \\ & P(C0004623) + \\ & P(C0036117) + \\ & P(C0038395) + \\ & P(C0039614) + \\ & P(C0036114) + \\ & P(C0041466) + \\ & P(C0021099) + \\ & P(C0032269) \end{aligned}$$



Example: probability of hand

$$IC(\text{Bacterial Infection}) = \log(P(\text{Bacterial Infection}))$$



IC-based measures: default

```
umls-similarity.pl tetanus salmonella  
-measure lin
```

- IC(concept) comes from UMLSonMedline
 - National Library of Medicine
 - Consists of concepts from 2009 AB UMLS and the frequency they occurred in medline using the Essie Search Engine (Ide et al 2007).
 - Medline: database of citations of biomedical and clinical articles.

Create your own IC file

- Two programs in utils/ directory
 - create-icfrequency.pl
 - create-icpropagation.pl

Create your own IC file: step through

RAW TEXT

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

Create your own IC file: step through

RAW TEXT

create-icfrequency.pl

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C00000000<>0
C0000039<>9594
C0000052<>1518
C0000097<>15978
C0000102<>2149
C0000163<>1078
```

...

Create your own IC file: step through

RAW TEXT

create-icfrequency.pl

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C00000000<>0
C0000039<>9594
C0000052<>1518
C0000097<>15978
C0000102<>2149
C0000163<>1078
```

...

Create your own IC file: step through

RAW TEXT

create-icfrequency.pl
--metamap

OPTIONS:

- METAMAP

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

SAB :: include SNOMEDCT
REL :: include PAR, CHD

C0000000<>0
C0000039<>9594
C0000052<>1518
C0000097<>15978
C0000102<>2149
C0000163<>1078

...

Create your own IC file: step through

RAW TEXT

create-icfrequency.pl
--term

OPTIONS:

- METAMAP
- SIMPLE STRING MATCH OF TERMS TO THE MRCONSO TABLE IN THE UMLS

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

SAB :: include SNOMEDCT
REL :: include PAR, CHD

C0000000<>0
C0000039<>9594
C0000052<>1518
C0000097<>15978
C0000102<>2149
C0000163<>1078

...

Create your own IC file: step through

RAW TEXT

create-icfrequency.pl

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not Be reliable because ...

SAB :: include SNOMEDCT
REL :: include PAR, CHD

C00000000<>0
C0000039<>9594
C0000052<>1518
C0000097<>15978
C0000102<>2149
C0000163<>1078
...

**MORE
ACCURATE**

OPTIONS:

- METAMAP
- SIMPLE STRING MATCH OF TERMS TO THE MRCONSO TABLE IN THE UMLS

FASTER

Create your own IC file: step through

RAW TEXT

create-icfrequency.pl

create-icpropagation.pl

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

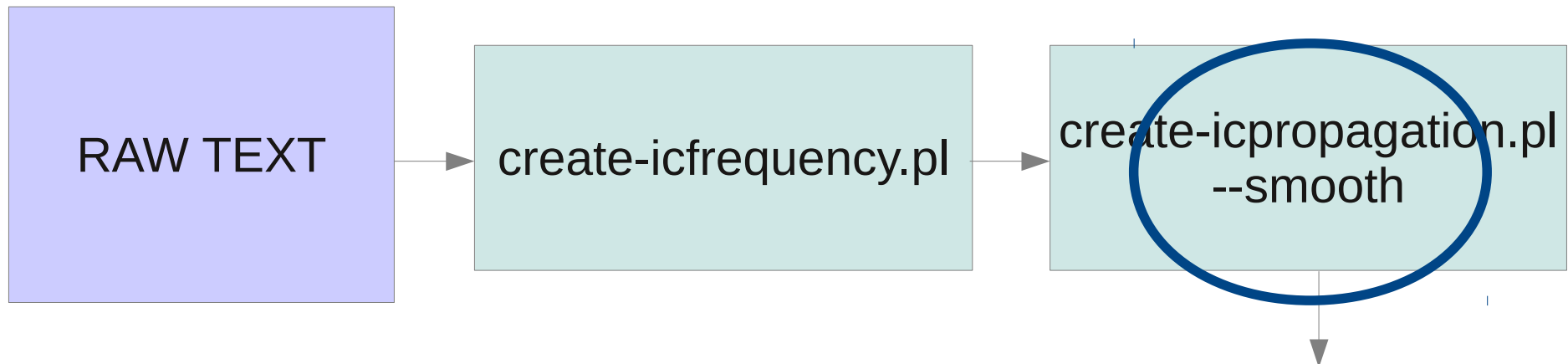
```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C00000000<>0
C0000039<>9594
C0000052<>1518
C0000097<>15978
C0000102<>2149
C0000163<>1078
...
```

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C00000000<>0
C0000039<>2.3e^06
C0000052<>3.21e^07
C0000097<>3.38e^06
C0000102<>9.12e^07
C0000163<>4.78e^05
...
```

Create your own IC file: step through



SMOOTHING OPTION:

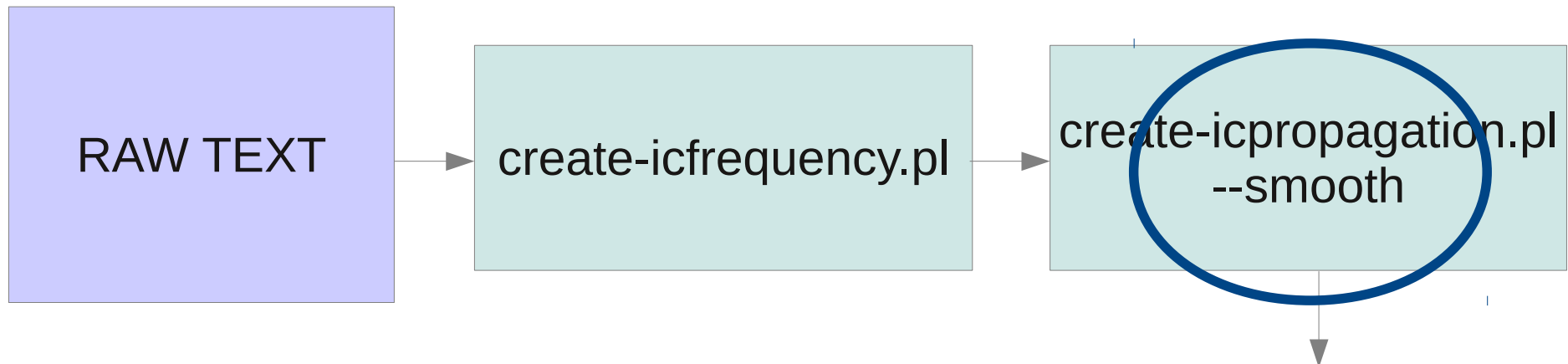
Laplace Add-1
Smoothing

Add 1 to each concept
in the taxonomy

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C0000000<>1
C0000039<>2.3e^06
C0000052<>3.21e^07
C0000097<>3.38e^06
C0000102<>9.12e^07
C0000163<>4.78e^05
...
```

Create your own IC file: step through



SMOOTHING OPTION:

Laplace Add-1
Smoothing

Rule thumb:

small file = smooth
large file != smooth

Add 1 to each concept
in the taxonomy

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C0000000<>1
C0000039<>2.3e^06
C0000052<>3.21e^07
C0000097<>3.38e^06
C0000102<>9.12e^07
C0000163<>4.78e^05
...
```


icfrequency and icpropagation

- UMLS::Similarity utils/directory
 - create-icfrequency.pl
 - --term option (default)
 - --metamap option
 - create-icpropagation.pl
 - --smooth

create-icfrequency.pl example

```
create-icfrequency.pl icfreq  
text  
--config config  
--metamap
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

RAW TEXT : text

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

create-icfrequency.pl example

```
create-icfrequency.pl icfreq  
text  
--config config  
--metamap
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

RAW TEXT : text

Background. The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

create-icfrequency.pl example

```
create-icfrequency.pl icfreq  
text  
--config config  
--metamap
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

RAW TEXT : text

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

create-icfrequency.pl example

```
create-icfrequency.pl icfreq  
text  
--config config  
--metamap
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

RAW TEXT : text

Background: The optimal femorotibial angle (FTA) after high tibial osteotomy (HTO) is still controversial. Our hypothesis was that FTA itself may not be reliable because ...

create-icfrequency.pl example

```
create-icfrequency.pl icfreq  
text
```

```
--config config  
--metamap
```

ICFREQUENCY FILE : icfreq

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

```
C00000000<>0  
C0000039<>9594  
C0000052<>1518  
C0000097<>15978  
C0000102<>2149  
C0000163<>1078
```

```
...
```

create-icpropagation.pl example

```
create-icpropagation.pl icprop  
icfreq  
--config config
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

ICFREQUENCY FILE : icfreq

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

```
C0000000<>0  
C0000039<>9594  
C0000052<>1518  
C0000097<>15978  
C0000102<>2149  
C0000163<>1078
```

```
...
```

create-icpropagation.pl example

```
create-icpropagation.pl icprop  
icfreq  
--config config
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

ICFREQUENCY FILE : icfreq

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

```
C0000000<>0  
C0000039<>9594  
C0000052<>1518  
C0000097<>15978  
C0000102<>2149  
C0000163<>1078
```

```
...
```


create-icpropagation.pl example

```
create-icfrequency.pl icprop  
icfreq  
--config config
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

ICFREQUENCY FILE : icfreq

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

```
C0000000<>0  
C0000039<>9594  
C0000052<>1518  
C0000097<>15978  
C0000102<>2149  
C0000163<>1078
```

...

create-icpropagation.pl example

```
create-icfrequency.pl icprop  
icfreq  
--config config
```

~~CONFIG FILE~~ config

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

ICFREQUENCY FILE : icfreq

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

```
C0000000<>0  
C0000039<>9594  
C0000052<>1518  
C0000097<>15978  
C0000102<>2149  
C0000163<>1078
```

...

create-icpropagation.pl example

```
create-icfrequency.pl icprop  
                    icfreq  
                    --config config
```

ICPROPAGATION FILE : icprop

```
SAB :: include SNOMEDCT  
REL :: include PAR, CHD
```

```
C0000000<>1  
C0000039<>2.3e^06  
C0000052<>3.21e^07  
C0000097<>3.38e^06  
C0000102<>9.12e^07  
C0000163<>4.78e^05
```

```
...
```

Using icpropagation file

```
umls-similarity.pl tetanus salmonella
    – config config
    – icpropagation icprop
    – measure lin
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT
REL:: include PAR, CHD
```

ICPROPAGATION FILE : icprop

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C0000000<>1
C0000039<>2.3e^06
C0000052<>3.21e^07
C0000097<>3.38e^06
C0000102<>9.12e^07
C0000163<>4.78e^05
```

```
...
```

Using icpropagation file

```
umls-similarity.pl tetanus salmonella
    – config config
    – icpropagation icprop
    – measure lin
```

CONFIG FILE: config

```
SAB :: include SNOMEDCT
REL:: include PAR, CHD
```

ICPROPAGATION FILE : icprop

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C0000000<>0
C0000039<>2.3e^06
C0000052<>3.21e^07
C0000097<>3.38e^06
C0000102<>9.12e^07
C0000163<>4.78e^05
```

```
...
```

Using icproapagation file

```
umls-similarity.pl tetanus salmonella  
    – config config  
    – icpropagation icprop  
    – measure lin
```

```
0.3636<>tetanus(C0039614)<>salmonella(C0036111)
```

Similarity overview

Similarity overview

CONFIG FILE

SAB :: include SNOMEDCT
REL:: include PAR, CHD

- umls-similarity.pl program
 - General options:
 - --config FILE

Similarity overview

CONFIG FILE

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

- umls-similarity.pl program
 - General options:
 - --config FILE
 - --measure MEASURE

MEASURES

```
path  
wup  
lch  
zhong  
nam  
lin  
res  
Jcn  
lesk  
vector
```

Similarity overview

CONFIG FILE

```
SAB :: include SNOMEDCT  
REL:: include PAR, CHD
```

- umls-similarity.pl program
 - General options:
 - --config FILE
 - --measure MEASURE
 - IC options:

MEASURES

```
path  
wup  
lch  
zhong  
nam  
lin  
res  
Jcn  
lesk  
vector
```

Similarity overview

CONFIG FILE

```
SAB :: include SNOMEDCT
REL:: include PAR, CHD
```

- umls-similarity.pl program

- General options:

- --config FILE
- --measure MEASURE

- IC options:

ICFREQUENCY

- --icfrequency FILE

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C00000000<>0
C0000039<>9594
C0000052<>1518
C0000097<>15978
C0000102<>2149
C0000163<>1078
...
```

MEASURES

```
path
wup
lch
zhong
nam
lin
res
Jcn
lesk
vector
```

Similarity overview

CONFIG FILE

```
SAB :: include SNOMEDCT
REL:: include PAR, CHD
```

- umls-similarity.pl program

- General options:

- --config FILE
- --measure MEASURE

- IC options:

ICFREQUENCY

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C00000000<>0
C0000039<>9594
C0000052<>1518
C0000097<>15978
C0000102<>2149
C0000163<>1078
...
```

- --icfrequency FILE
- --icpropagation FILE

ICPROPAGATION

```
SAB :: include SNOMEDCT
REL :: include PAR, CHD
```

```
C00000000<>1
C0000039<>2.03e^06
C0000052<>3.21e^07
C0000097<>3.38e^06
C0000102<>0.12e^07
C0000163<>1.07e^06
...
```



MEASURES

```
path
wup
lch
zhong
nam
lin
res
Jcn
lesk
vector
```

UML::Similarity web interface

- Located:
 - <http://atlas.ahc.umn.edu>
 - <http://maraca.d.umn.edu>

UMLS::Similarity web interface

Applications Places System   ?

Similarity Google Docs - Home weekly: 2012 January Google News Twitter / Home

atlas.ahc.umn.edu/cgi-bin/uMLS_similarity.cgi

Google Gmail Research St. Paul Public Lib... Twitter My Yahoo! Other Bookmarks

UMLS::Similarity Web Interface

UMLS::Similarity is a freely available open source software package that can be used to obtain the similarity or relatedness between two biomedical terms from the [Unified Medical Language System \(UMLS\)](#).

DIRECTIONS: You may enter any two terms or [Concept Unique Identifiers \(CUIs\)](#) below. If terms are entered, then the relatedness or similarity of the possible CUIs will be computed and the pair with the highest score returned. [The difference between similarity and relatedness is](#)

[Detailed instructions.](#)
[About the Similarity Measures.](#)
[About the Relatedness Measures.](#)

Term 1:
Term 2:

Semantic Similarity

SAB:
REL:
Similarity:

Semantic Relatedness

SABDEF:
RELDEF:
Relatedness:

Similarity - Google Chro... ihi 2012 tutorial - btmci...

UMLS::Similarity web interface

UMLS::Similarity Web Interface

UMLS::Similarity is a freely available open source software package that can be used to obtain the similarity or relatedness between two biomedical terms from the [Unified Medical Language System \(UMLS\)](#).

DIRECTIONS: You may enter any two terms or [Concept Unique Identifiers \(CUIs\)](#) below. If terms are entered, then the relatedness or similarity of the possible CUIs will be computed and the pair with the highest score returned. [The difference between similarity and relatedness is](#)

[Detailed instructions.](#)
[About the Similarity Measures.](#)
[About the Relatedness Measures.](#)

Term 1:
Term 2:

Semantic Similarity

SAB:
REL:
Similarity:



CLICK FOR SIMILARITY

Semantic Relatedness

SABDEF:
RELDEF:
Relatedness:

CLICK FOR RELATEDNESS

UML::Similarity web interface

Applications Places System   ?

Similarity Google Docs - Home weekly: 2012 January Google News Twitter / Home

atlas.ahc.umn.edu/cgi-bin/u/mls_similarity.cgi?word1=tetanus&word2=salmonella&sab=MSH&rel=PAR%2FCHD&similarity=path&button=Compute+Similarity&sabdef=UMLS_ALL&reldef=CUI%2FPAR%2FCHD%2FRB%2FRN&relatedness=

Google Gmail Research St. Paul Public Lib... Twitter My Yahoo! Other Bookmarks

UMLS::Similarity Web Interface

UMLS::Similarity is a freely available open source software package that can be used to obtain the similarity or relatedness between two biomedical terms from the [Unified Medical Language System](#) (UMLS).

[View Definitions](#)

[View Shortest Path](#)

Results:

The similarity of tetanus (C0039614) and salmonella (C0036111) using Path Length (path) is 0.0769.

Using:

SAB :: include MSH

REL :: include PAR/CHD

[View relatedness of all possible senses](#)

DIRECTIONS: You may enter any two terms or [Concept Unique Identifiers](#) (CUIs) below. If terms are entered, then the relatedness or similarity of the possible CUIs will be computed and the pair with the highest score returned. [The difference between similarity and relatedness is](#)

[Detailed instructions.](#)

[About the Similarity Measures.](#)

[About the Relatedness Measures.](#)

Term 1:

Term 2:

Semantic Similarity

SAB:

REL:

UMLS: Similarity web interface

The screenshot displays the UMLS Similarity web interface in a Google Chrome browser. The browser's address bar shows the URL: `atlas.ahc.umn.edu/cgi-bin/umls_similarity.cgi?word1=tetanus&word2=salmonella&sab=MSH&rel=PAR%2FCHD&similarity=path`. The page title is "UMLS::Similarity Web Interface".

UMLS::Similarity is a freely available open source software package that can be used to obtain the similarity or relatedness between two biomedical terms from the [Unified Medical Language System \(UMLS\)](#).

View Definitions ← **CLICK**

View Shortest Path

Results:

The similarity of tetanus ([C0039614](#)) and salmonella ([C0036111](#)) using Path Length (path) is 0.0769.

Using:

SAB :: include MSH

REL :: include PAR/CHD

[View relatedness of all possible senses](#)

DIRECTIONS: You may enter any two terms or [Concept Unique Identifiers \(CUIs\)](#) below. If terms are entered, then the relatedness is returned. [The difference between similarity and relatedness is](#)

[Detailed instructions.](#)
[About the Similarity Measures.](#)
[About the Relatedness Measures.](#)

Term 1:

Term 2:

Semantic Similarity

SAB:

REL:

Definition:

tetanus (C0039614)

MSH : A disease caused by tetanospasmin, a powerful protein toxin produced by CLOSTRIDIUM TETANI. Tetanus usually occurs after an acute injury, such as a puncture wound or laceration. Generalized tetanus, the most common form, is characterized by tetanic muscular contractions and hyperreflexia. Localized tetanus presents itself as a mild condition with manifestations restricted to muscles near the wound. It may progress to the generalized form.

CSP : disease caused by tetanospasmin, a powerful protein toxin produced by Clostridium tetani; tetanus usually occurs after an acute injury, such as a puncture wound or laceration; generalized tetanus, the most common form, is characterized by tetanic muscular contractions and hyperreflexia; localized tetanus presents itself as a mild condition with manifestations restricted to muscles near the wound.

MEDLINEPLUS :

Tetanus is a serious illness caused by tetanus bacteria. The bacteria live in soil, saliva, dust and manure. The bacteria usually enter the body through a deep cut, like those you might get from cutting yourself with a knife or stepping on a nail.

The infection causes painful tightening of the muscles, usually all over the body. It can lead to "locking" of the jaw, which makes it impossible to open your mouth or swallow. If this happens, you could die of suffocation.

UMLS::Similarity web interface

The screenshot displays the UMLS::Similarity web interface in a Google Chrome browser. The browser's address bar shows the URL: `atlas.ahc.umn.edu/cgi-bin/umls_similarity.cgi?word1=tetanus&word2=salmonella&sab=MSH&rel=PAR%2FCHD&similarity=path`. The page title is "UMLS::Similarity Web Interface".

The main content area includes a description of the software: "UMLS::Similarity is a freely available open source software package that can be used to obtain the similarity or relatedness between two biomedical terms from the Unified Medical Language System (UMLS)." Below this, there are links for "View Definitions" and "View Shortest Path". A blue arrow points to the "View Shortest Path" link, with the word "CLICK" next to it.

The "Results:" section shows the similarity of tetanus (C0039614) and salmonella (C0036111) using Path Length (path) is 0.0769. It also lists the settings: "SAB :: include MSH" and "REL :: include PAR/CHD". There are links for "View relatedness of all possible senses", "Detailed instructions", "About the Similarity Measures", and "About the Relatedness Measures".

At the bottom, there are input fields for "Term 1:" (tetanus) and "Term 2:" (salmonella). Below these are dropdown menus for "SAB:" (MSH) and "REL:" (PAR/CHD).

The right sidebar contains a section titled "Shortest Path Information" which states: "The shortest path between C0039614 (Tetanus) and C0036111 (salmonellas) is: C0039614 (Tetanus) => C0009062 (Infections, Clostridium) => C0085426 (Infections, Gram-Positive Bacterial) => C0004623 (Infections, Bacterial) => C0004615 (Bacterial Infections and Mycoses) => C0012674 (Diseases (MeSH Category)) => C1256741 (Topical Descriptor) => C1256748 (Organisms (MeSH Category)) => C0004611 (Bacteria.) => C0018150 (bacterium gram-negative) => C0018152 (Gram Negative Facultatively Anaerobic Rods) => C0014346 (Enterobacteriaceae) => C0036111 (salmonellas)".

At the bottom right of the sidebar, there is a visitor counter showing "845 Visitors" and the date range "5 Apr 2011 - 13 Jan 2012".

UMLS::Similarity web interface

Applications Places System Sat Jan 14, 10:17 AM bridget

Similarity Google Docs - Home weekly: 2012 January Google News How to Take a Screensh

atlas.ahc.umn.edu/cgi-bin/umls_similarity.cgi

Google Gmail Research St. Paul Public Lib... Twitter My Yahoo! Other Bookmarks

UMLS::Similarity Web Interface

UMLS::Similarity is a freely available open source software package that can be used to obtain the similarity or relatedness between two biomedical terms from the [Unified Medical Language System \(UMLS\)](#).

DIRECTIONS: You may enter any two terms or [Concept Unique Identifiers \(CUIs\)](#) below. If terms are entered, then the relatedness or similarity of the possible CUIs will be computed and the pair with the highest score returned. [The difference between similarity and relatedness is](#)

[Detailed instructions.](#)
[About the Similarity Measures.](#)
[About the Relatedness Measures.](#)

Term 1:

Term 2:

Semantic Similarity

SAB:

REL:

Similarity:

MODIFY SOURCES, RELATIONS AND SIMILARITY MEASURE

Semantic Relatedness

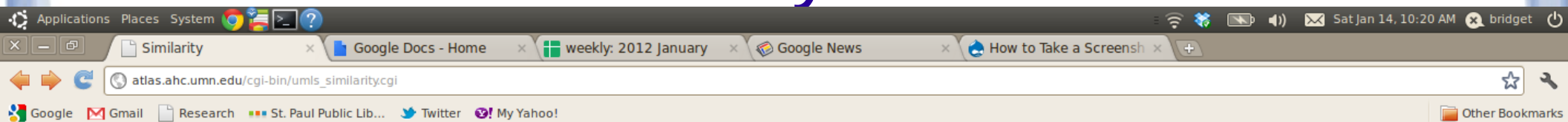
SABDEF:

RELDEF:

Relatedness:

Similarity - Google Chro...

UMLS::Similarity web interface



UMLS::Similarity Web Interface

UMLS::Similarity is a freely available open source software package that can be used to obtain the similarity or relatedness between two biomedical terms from the [Unified Medical Language System \(UMLS\)](#).

DIRECTIONS: You may enter any two terms or [Concept Unique Identifiers \(CUIs\)](#) below. If terms are entered, then the relatedness or similarity of the possible CUIs will be computed and the pair with the highest score returned. [The difference between similarity and relatedness is](#)

[Detailed instructions.](#)
[About the Similarity Measures.](#)
[About the Relatedness Measures.](#)

Term 1:
Term 2:

Semantic Similarity

SAB:
REL:
Similarity:

Semantic Relatedness

SABDEF:
RELDEF:
Relatedness:

**MODIFY SOURCES, RELATIONS AND
RELATEDNESS MEASURE**

Thank you

Questions?

Relatedness Measures Contained in the UMLS-Similarity Package

Measuring the Similarity and
Relatedness of Concepts in the
Medical Domain : IHI 2012 Tutorial

Ying Liu

College of Pharmacy
University of Minnesota
Minneapolis, MN

liux0395@umn.edu
<http://www.tc.umn.edu/~liux0395>

Outline

- Introduction of the semantic relatedness measures: lesk and vector
- How to use UMLS-Similarity to get the relatedness score

Ontology dependent and independent measures

- **Ontology dependent measures**
 - relay on the concept hierarchies or ontologies
 - is-a, has-part, and is-a-part-of...
 - path based: path, Wu&Palmer, and Leacock&Chodorow...
 - Information content (IC) based: Resnik, Lin and Jiang&Conrath
- **Ontology independent measures**
 - rely on the related concepts have a similar context
 - lesk : Adapted Lesk, Banerjee and Pedersen (2003)
 - gloss vector : Patwardhan and Pedersen (2006)

Lesk semantic relatedness measure

- Semantic relatedness is a function of the overlap between their definitions

$$rel_{lesk} = \sum_{i=1}^n freq_{overlap_i} * length_{overlap_i}^2$$

Example:

- influenza** : infectious disease, fever, muscle pains, and general discomfort
- aspirin** : relieve pains, reduce fever, an anti-inflammatory medication

$$rel_{lesk} = 1 * 1^2 + 1 * 1^2 = 2$$

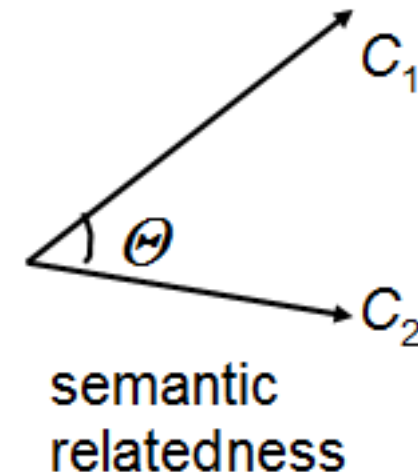
Disadvantages of lesk method

- Based strictly on definitions and doesn't use any other knowledge source
- A word could have several forms
example: Minnesota vs. MN
- Different words have the same semantic meaning
example: utility vs. usage

Vector semantic relatedness measure

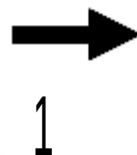
- Definition
 - **influenza** : infectious disease, fever, muscle pains, and general discomfort.
 - **aspirin** : relieve pains, reduce fever, an anti-inflammatory medication
- Co-occurrence vector

Infectious: bacterial diagnoses fungi
disease: behavior body cost feel research risks
fever: attack body case fell health
muscle: ache change exercise injury
pains: nausea headache
general: analysis appear body diet family
discomfort: anger felt nausea stress
relieve: chest drug time pains
reduce: abnormal access asthma clinical error
anti-inflammatory: drugs therapy
medication: abuse choice diet expert patient



The procedure of the second-order context vector semantic relatedness method

the original corpus for
building the co-occurrence
matrix.

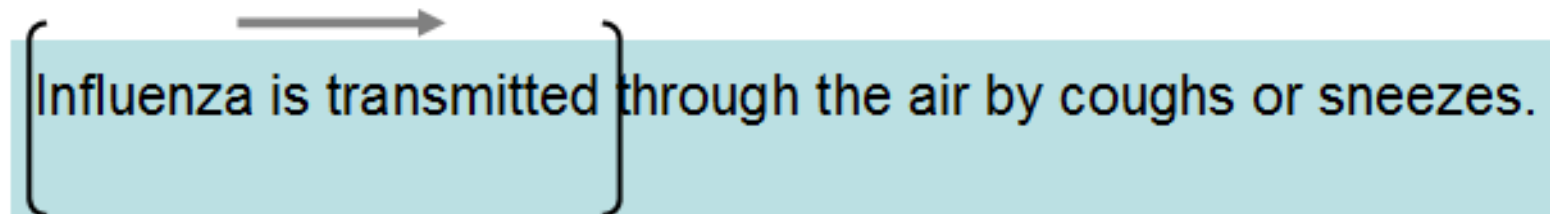


scan text and count
the bi-grams

```
...  
last<>year<>34  
long<>term<>100  
new<>york<>2  
on<>monday<>5  
...
```

Step1: A bi-gram example with window=3

move the window from left to right

Influenza is transmitted through the air by coughs or sneezes.

window=3

influenza<>is<>
is<>through<>
through<>the<>
the<>by<>
by<>coughs<>
or<>sneezes

is<>transmitted<>
transmitted<>through<>
through<>air<>
air<>by<>
by<>or<>
coughs<>sneezes

influenza<>transmitted<>
transmitted<>the<>
the<>air<>
air<>coughs<>
coughs<>or<>

Step1: A bi-gram example with window=3

move the window from left to right

Influenza is transmitted through the air by coughs or sneezes.

window=3

influenza<>is<>
is<>through<>
through<>the<>
the<>by<>
by<>coughs<>
or<>sneezes

is<>transmitted<>
transmitted<>through<>
through<>air<>
air<>by<>
by<>or<>
coughs<>sneezes

influenza<>transmitted<>
transmitted<>the<>
the<>air<>
air<>coughs<>
coughs<>or<>

Step1: how to get the bi-gram list

- Build the bi-gram list by Text-NSP
 - Ngram Statistics Package
 - download at : <http://sourceforge.net/projects/ngram/>
- count.pl or huge-count.pl
 - generate the bi-gram list
- count2huge.pl
 - if use count.pl, need to convert the bi-gram order by count2huge.pl

The procedure of the second-order context vector semantic relatedness method

the original corpus for building the co-occurrence matrix.



1

scan text and count the bi-grams

```
...  
last<>year<>34  
long<>term<>100  
new<>york<>2  
on<>monday<>5  
...
```

2

build the co-occurrence matrix.

first order context vector

	W_1	W_2	W_3	...	W_n
W_1	0	2	5	...	11
W_2	1	0	4	...	3
...					
W_m	6	1	0	...	9

$n \times m$ co-occurrence matrix

Sept 2: $n \times m$ co-occurrence matrix

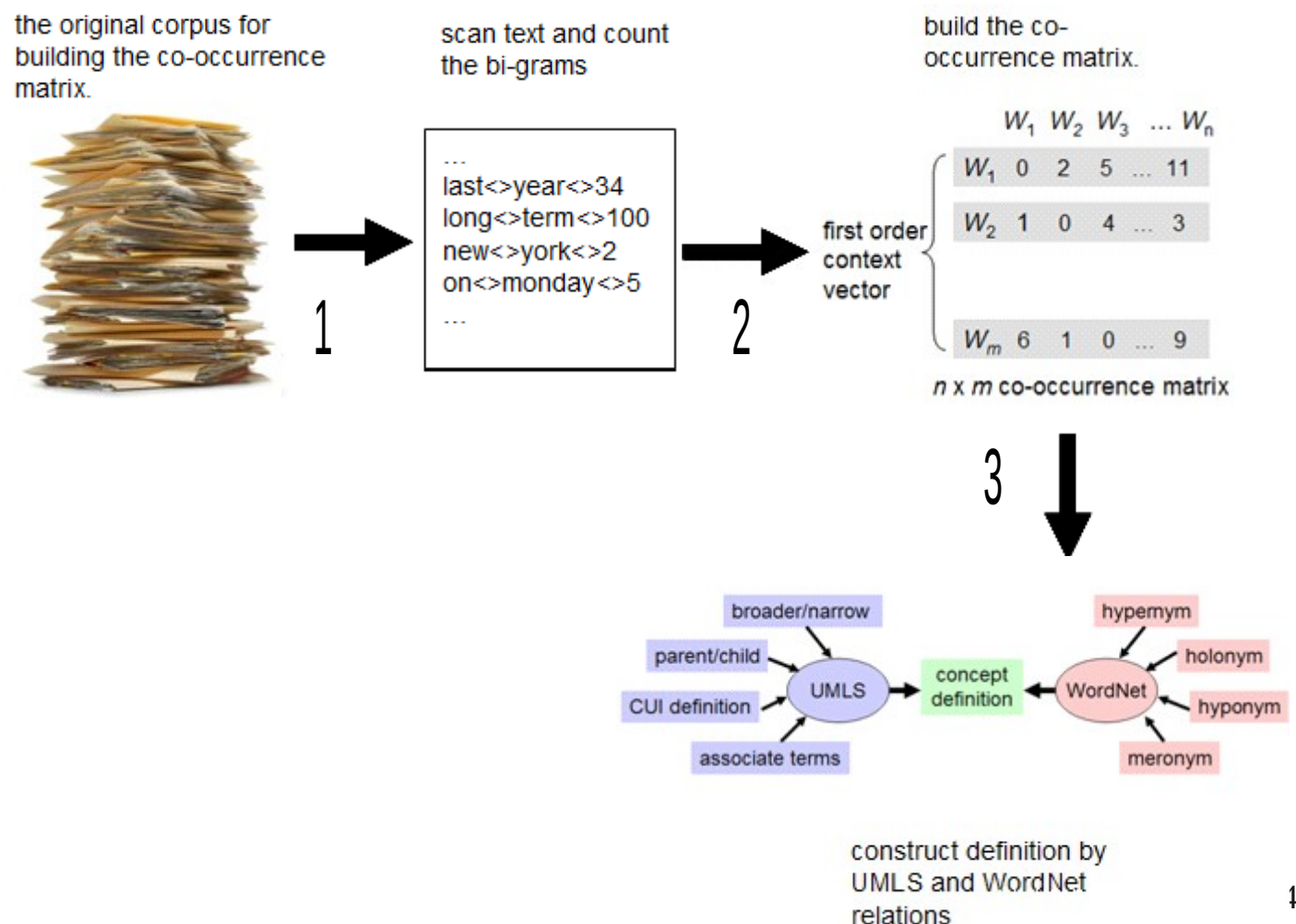
		W_1	W_2	W_3	...	W_n
first-order context vector	W_1	0	2	5	...	11
	W_2	1	0	4	...	3
	W_m	6	1	0	...	9

Infectious: bacterial 2 diagnoses 1 fungi 3
disease: behavior 1 body 5 cost 3 feel 7 research 5 risks 2
fever: attack 3 body 5 case 10 fell 6 health 8
muscle: ache 2 change 5 exercise 9 injury 2
pains: nausea 8 headache 20
general: analysis 5 appear 6 body 9 diet 4 family 7
discomfort: anger 2 fell 4 nausea 6 stress 3
relieve: chest 6 drug 4 time 7 pains 9
reduce: abnormal 2 access 5 asthma 6 clinical 9 error 3
anti-inflammatory: drugs 4 therapy 7
medication: abuse 3 choice 6 diet 8 expert 4 patient 3

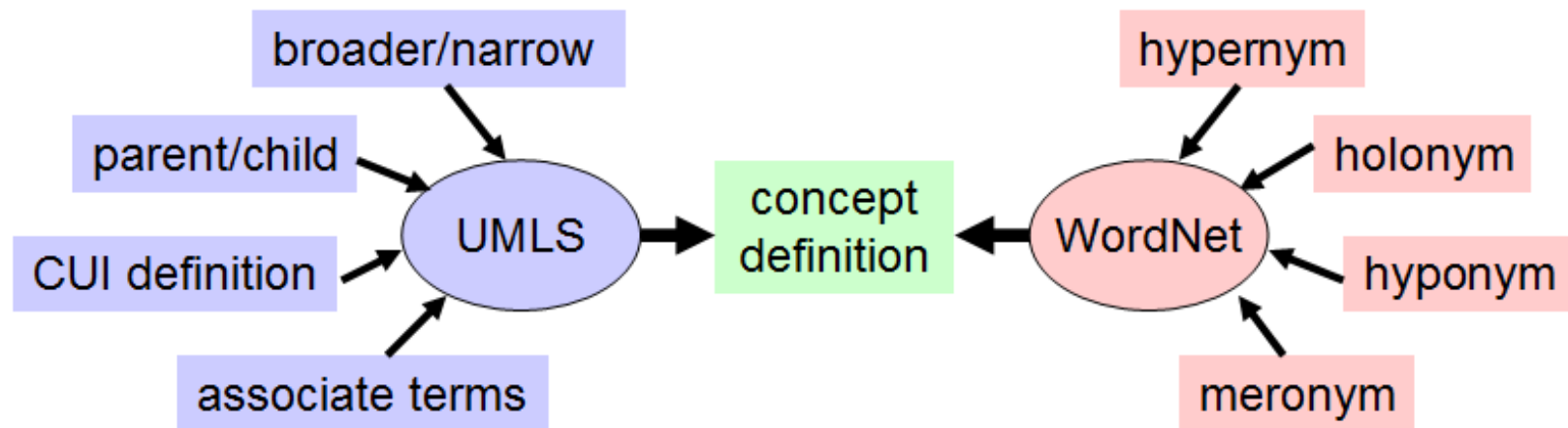
Sept 2: $n \times m$ co-occurrence matrix

- vector-input.pl of UMLS-Similarity
 - read the sorted bi-gram list
- co-occurrence matrix file has two parts
 - index file: record each vector's position and length
 - matrix file: record the vectors

The procedure of the second-order context vector semantic relatedness method



Step 3: Organization of a concept's definition



influenza : infectious disease, fever, muscle pains, and general discomfort.
(CUI)

+ **disease** (parent) : an abnormal condition affecting the body of an organism.

+ **cold** (child) : running nose, sneeze.

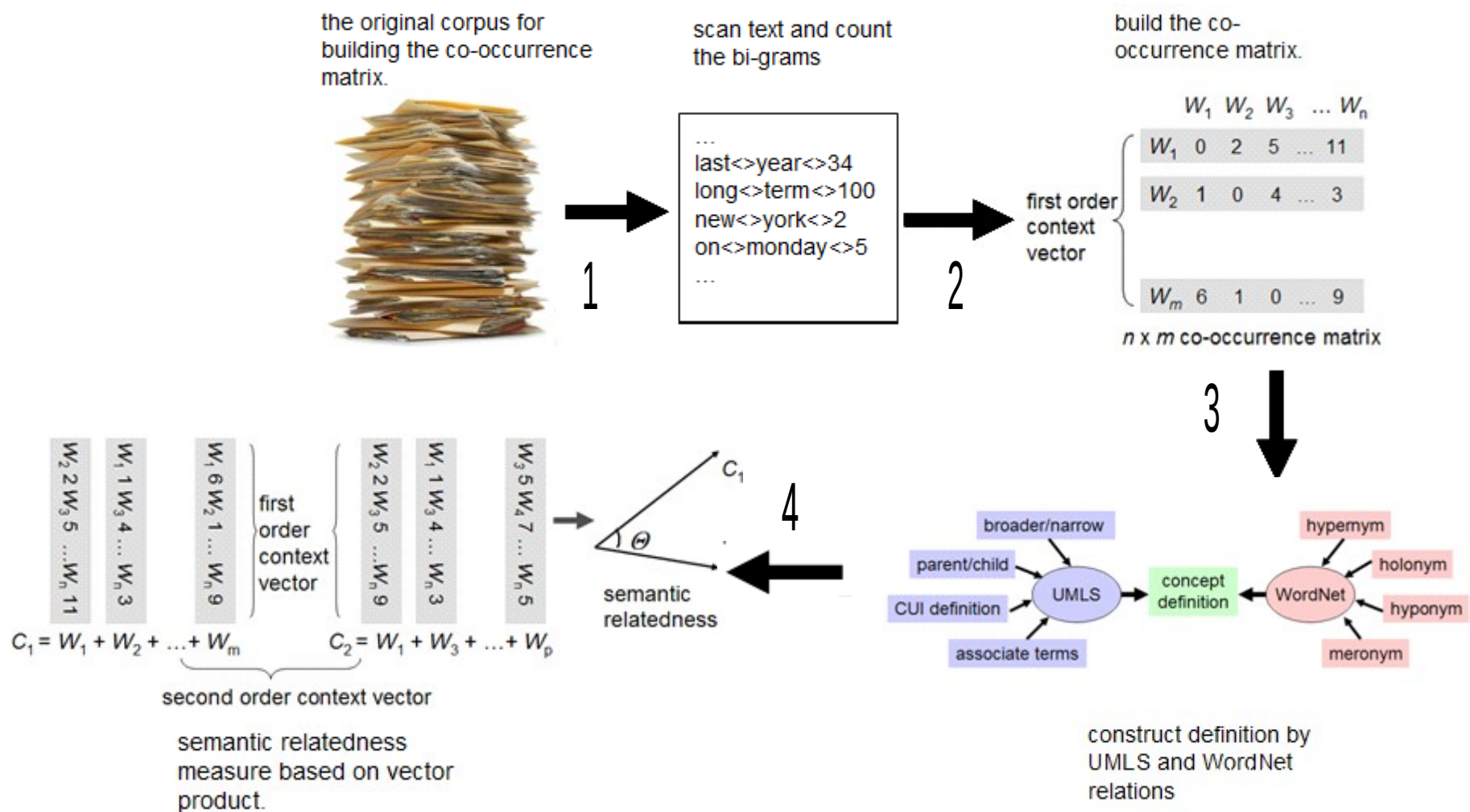
+ **cough** (associate terms)

+ influenza is transmitted through the air by coughs and sneezes. (WordNet)

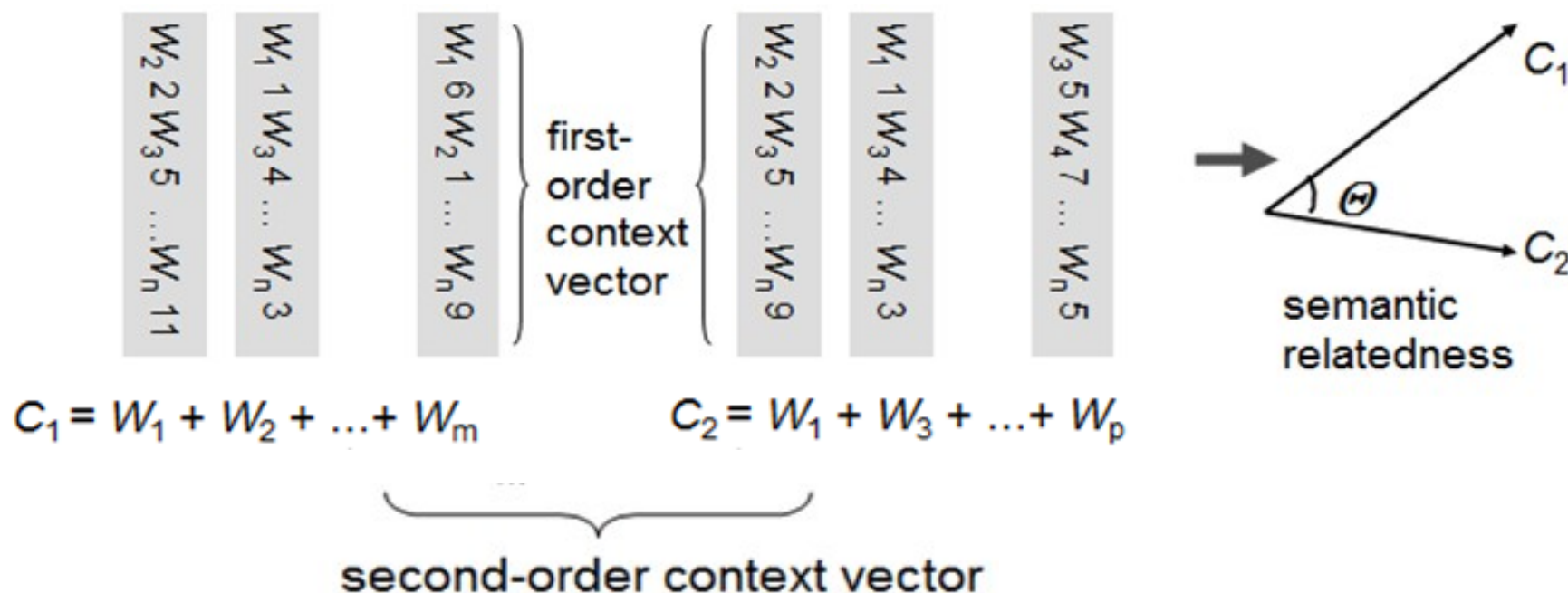
Step 3: Organization of a concept's definition

- --config option to define source and relations
- --dictfile option to import external definitions from WordNet or other sources

The procedure of the second-order context vector semantic relatedness method



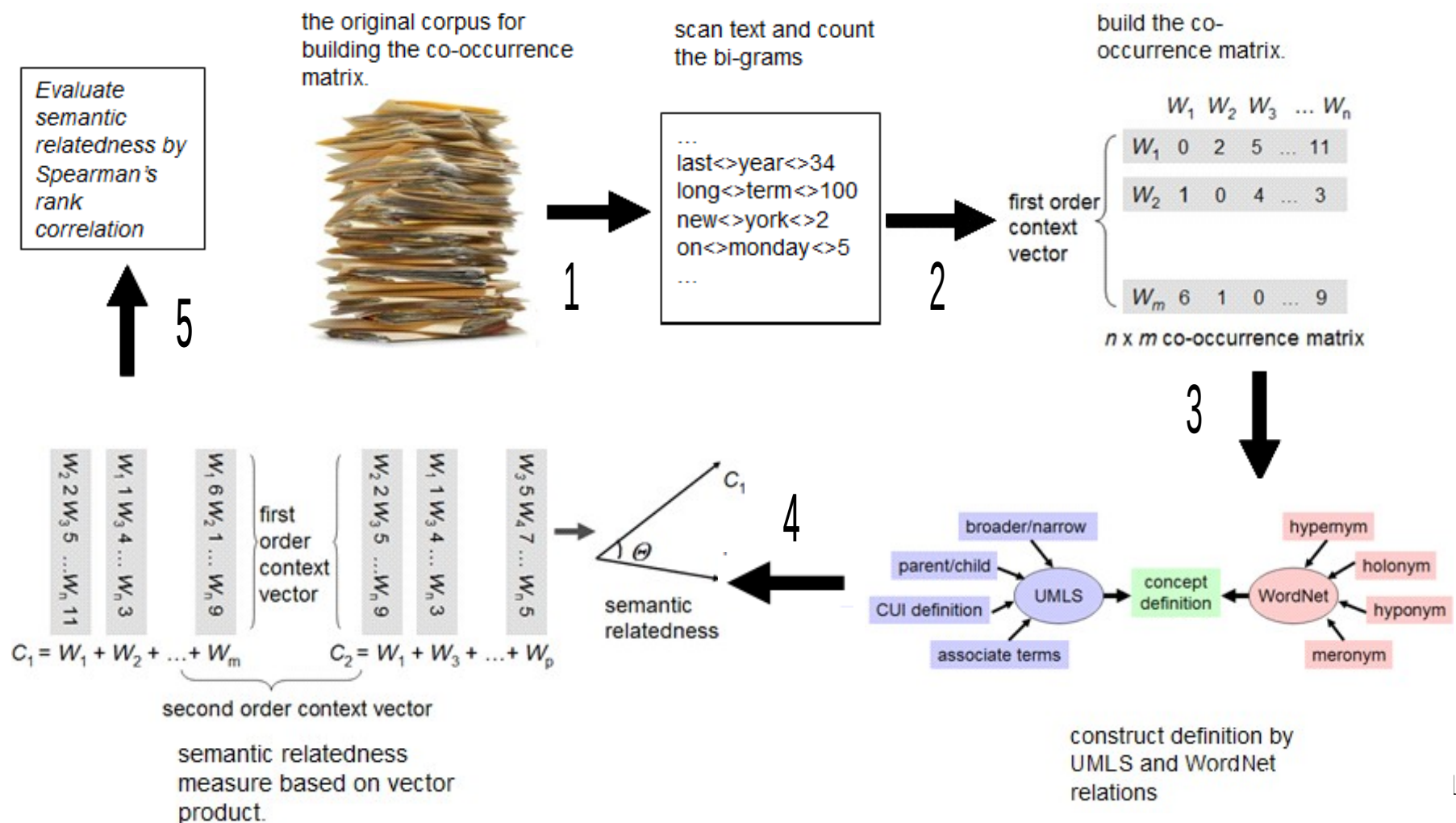
Step 4: Geometric explanation of the semantic relatedness measure based on vector product



Influenza (C_1) : bacterial 2 diagnoses 1 fungi 3 behavior 1 body 19 cost 3 feel 7 research 5 risks 2 attack 3 case 10 fell 6 health 8 ache 2 change 5 exercise 9 injury 2 nausea 8 headache 20 analysis 5 appear 6 diet 4 family 7 anger 2 fell 4 nausea 6 stress 3

aspirin (C_2): chest 6 drug 4 time 7 pains 9 nausea 8 headache 20 abnormal 2 access 5 asthma 6 clinical 9 error 3 attack 3 body 5 case 10 fell 6 health 8 drugs 4 therapy 7 abuse 3 choice 6 diet 8 expert 4 patient 3

The procedure of the second-order context vector semantic relatedness method



Command to get the vector relatedness

- `umls-similarity.pl --measure vector --vectorindex index_file --vectormatrix matrix_file influenza aspirin`
- Other options
 - `--config`
 - `--dictfile`
 - `--compoundfile`
 - `--doubledef`
 - `--stoplist`
 - `--defraw`
 - `--stem`
 - `--debugfile`

http://atlas.ahc.umn.edu/cgi-bin/umls_similarity.cgi

☐

UMLS::Similarity Web Interface

UMLS::Similarity is a freely available open source software package that can be used to obtain the similarity or relatedness between two biomedical terms from the [Unified Medical Language System \(UMLS\)](#). Please note, the link to the UMLS::Similarity package is severed at this time for the purpose of anonymity.

[View Definitions](#)

Results: The relatedness of influenza (C0021400) and aspirin (C0004057) using Vector Measure (vector) is 0.2601.

The relatedness of influenza (C0021400) and aspirin (C0004057) using Vector Measure (vector) is 0.2601.

Using:

SABDEF:: include UMLS_ALL

RELDEF:: include CUI/PAR/CHD/RB/RN

[View relatedness of all possible senses](#)

DIRECTIONS: You may enter any two terms or [Concept Unique Identifiers \(CUIs\)](#) below. If terms are entered, then the relatedness or similarity of the possible CUIs will be computed and the pair with the highest score returned. [The difference between similarity and relatedness is ...](#)

[Detailed instructions](#)

[About the Similarity Measures](#)

[About the Relatedness Measures](#)

Term 1: influenza

Term 2: aspirin

Term 1: influenza

Term 2: aspirin

Semantic Similarity

SAB: MSH

REL: PAR/CHD

Similarity: Path Length (path)

Semantic Relatedness

SABDEF: UMLS_ALL

RELDEF: CUI/PAR/CHD/RB/RN

[Show version info](#)

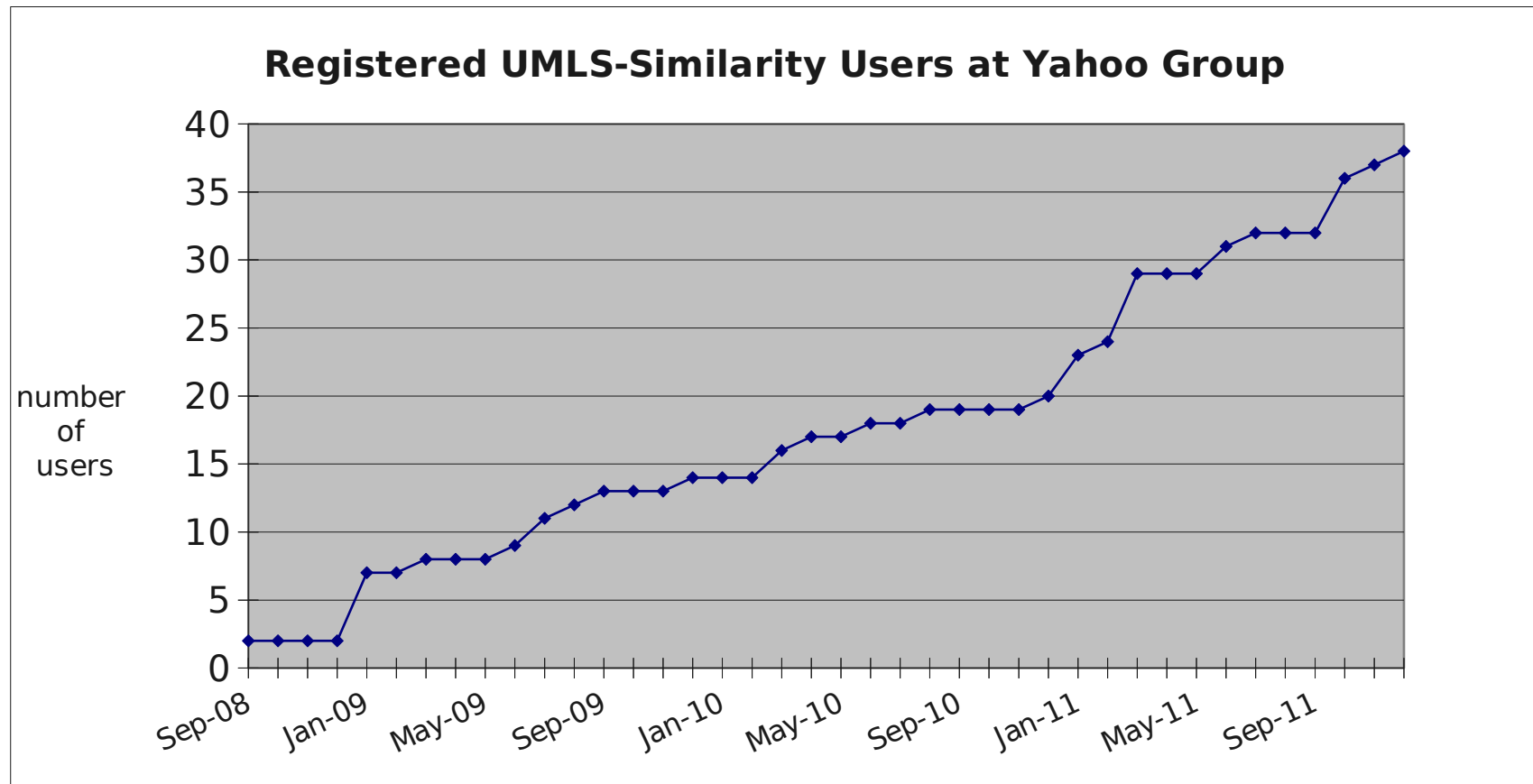
This interface is based on the [WordNet::Similarity web interface](#)

1

Applications of UMLS-Similarity

- Used UMLS-Similarity to calculate the gene and disease similarity. “Finding Disease Similarity Based on Implicit Semantic Similarity” at Journal of Biomedical Informatics 2011 by Mathur and Dinakarpandian
- Used UMLS-Similarity to connect relevant users together in the conversation and also provide contextual recommendations relevant to the health information conversation system Cobot. “Socio-Semantic Health Information Access” at Association for the Advancement of Artificial Intelligence 2011 by Sahay and Ram
- Used UMLS-Similarity to improve the performance of the classifier OWCP (one word conjunct pairs). “Coordination Resolution in Biomedical Texts” Ph. D dissertation 2011 by Philip Ogren

The increase of registered UMLS-Similarity Users at Yahoo group



Sign up at : <http://tech.groups.yahoo.com/group/umls-similarity>

Anonymous users of the web interface are from
46 countries and territories

http://atlas.ahc.umn.edu/cgi-bin/uuls_similarity.cgi



Summary

- Second order co-occurrence vector semantic relatedness method
 - Use proper relationship to construct the definition
 - Choice proper corpus to build the co-occurrence vector
 - Does not rely on the hierarchical structure
- http://atlas.ahc.umn.edu/cgi-bin/umls_similarity.cgi

Tomorrow...

- 10:30am-12:00pm Session D3-1A BioNLP (Vista Room)
- Semantic Relatedness Study Using Second Order Co-Occurrence Vector Computed by Biomedical Corpora, UMLS and WordNet
 - Ying Liu, Bridget T. McInnes, Ted Pedersen, Serguei Pakhomov and Genevieve Melton-Meaux

Thank You !

References

- Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th Annual International Conference on Systems Documentation, 1986;24–26.
- Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In Proceedings of the EACL workshop, Making sense of sense: bringing computational linguistics and psycholinguistics together. 2006;1-8.
- Patwardhan S. Incorporating dictionary and corpus information into context vector measure of semantic relatedness. Master of science Thesis, Duluth, MN: Department of Computer Science. Duluth: University of Minnesota; 2003.
- Pedersen T, Pakhomov S, Patwardhan S, Chute CG. Measures of Semantic Similarity and Relatedness in the biomedical domain. Journal of Biomedical Informatics. 2007;40(3);287-99.
- PakhomovS, McInnesB, AdamT, LiuY, PedersenT, Melton G. Semantic similarity and relatedness between clinical terms: an experimental study. In Prceedings of AMIA. 2010;572-76.
- Pedersen T, Patwardhan S, Michelizzi J. 2004. WordNet:: Similarity: measuring the relatedness of concepts. In Demonstration Papers at HLT-NAACL. 2004;38–41.
- McInnes B, Pedersen T, Pakhomov S. UMLS-Interface and UMLS-Similarity: Open Source Software for measuring paths and semantic similarity. In Proceedings of the Annual Symposium of the American Medical Informatics Association. 2009;431-35.
- Wu Z, Palmer, M. Verbs semantics and lexical selection. In Proceedings of the 32nd Meeting of ssociation of Computational Linguistics. 1994;133–38.
- Leacock C, Chodorow M. Combining local context and WordNet similarity for word sense identification. In WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA. 1998;265–83.

Evaluating and deploying measures of semantic relatedness

Serguei VS Pakhomov

College of Pharmacy
University of Minnesota
Minneapolis, MN

pakh0002 at umn dot edu

Outline

- Different types of evaluation (direct vs indirect)
- Creating a new reference standard
- Using existing resources
- Available reference standards (M&C, R&G, MayoSRS, MiniMayoSRS, UMNSRS)
- Evaluation metrics and statistical considerations (correlation, precision/recall, inter-rater agreement)
- Application to WSD

Evaluation

- Direct vs indirect approaches
 - Direct – evaluation against a manually annotated corpus of word/term pairs
 - Indirect – evaluation using measures to accomplish another task (e.g., WSD)

Evaluation

- Direct vs. Indirect approaches
 - Direct – evaluation against a manually annotated corpus of word/term pairs
 - Indirect – evaluation using measures to accomplish another task (e.g. WSD)

Pros and Cons

- Direct approaches
 - Pros:
 - relatively easy to implement and interpret
 - eliminate potential confounders present in indirect approaches
 - allow some control over the definition of relatedness/similarity
 - enable easy debugging and fine-tuning
 - Cons:
 - limited inter-annotator agreement due to individual variability
 - limited generalizability

Pros and Cons

- Indirect Approaches

- Pros:

- assess the “real-world” impact of relatedness and similarity measures
 - less prone to reference standard reliability issues
 - Is not constrained by specific definition of relatedness/similarity – more generalizable

- Cons:

- relatively difficult to implement unless the dataset already exists (e.g. WSD)
 - requires a study design that can isolate the effects of measures from other task variables

Direct Evaluation

- Creating a new reference standard
 - 1. select the data (word pairs) to be manually assessed
 - 2. determine the response variable (discrete, vs. continuous, nominal vs. numeric)
 - 3. determine the mode of presentation (paper, computer, spreadsheet, web-based, etc.)
 - 4. select and train annotators
 - 5. annotate
 - 6. evaluate the reliability of the dataset

Direct Evaluation

- Using existing resources as a reference
 - Miller and Charles (General Eng)
 - Rubenstein and Goodenough (General Eng)
 - MayoSRS and MiniMayoSRS (Medical)
 - UMNSRS (Medical)

Indirect Evaluation

- Creating a new reference standard
 - Too many variations to discuss
 - In general, follow standard guidelines for the specific task and determine the reliability of the reference standard
 - Example: WSD
 - manually label (or use already labeled) corpus of disambiguated text with known reliability

Indirect Evaluation

- Using existing resources
 - NLM WSD dataset
 - <http://wsd.nlm.nih.gov/>
 - 50 frequent UMLS concepts, each with 100 ambiguous instances
 - 11 annotators (majority vote)
 - Kappa statistic is discussed but not reported
 - <http://wsd.nlm.nih.gov/info/AMIA2001-weeber.pdf>

Statistical Considerations

- Reliability of reference standards
 - Unadjusted for chance
 - Correlation
 - Percent agreement
 - Adjusted for chance
 - Kappa
 - Intra-class correlation coefficient

Available Reference Standards

- Miller and Charles (M&C) and Rubenstein and Goodenough (R&G)
 - General English but...
 - May be used for annotator training purposes
 - Rubenstein and Goodenough (1965)
 - Synonymy judgements
 - 65 noun pairs
 - 51 subjects
 - Scale 0.0 - 4.0
 - $R = 0.85$ (repeated measures 2 weeks apart)

Miller and Charles Set

- 30 pairs of English nouns from R&G set

Word1	Word2	Mean score
Car	Automobile	3.92
Gem	Jewel	3.84
Coast	Shore	3.70
Monk	Oracle	1.10

- High correlation with R&G set – $r = 0.97$
- Later reproduced by Resnik (1999) – $r = 0.90$

Available Reference Standards

- Mayo Semantic Relatedness Standard (SRS)
 - Initially generated by a rheumatologist
 - 101 **term** pairs (originally 120) from four categories
 - Closely related, somewhat related, somewhat unrelated, completely unrelated
 - Not restricted to single word concepts
 - Annotated by 3 physicians and 12 medical coders
 - $r = 0.51$

Mayo SRS Format

Mean	CUI1	CUI2	TERM1	TERM2
6.69	C0311394	C0231685	difficulty walking	antalgic gait
2.38	C0035450	C0034079	rheumatoid nodule	lung nodule
1.00	C0409162	C0333286	hand splint	splinter hemorrhage
1.00	C0011849	C0032584	diabetes	polyp

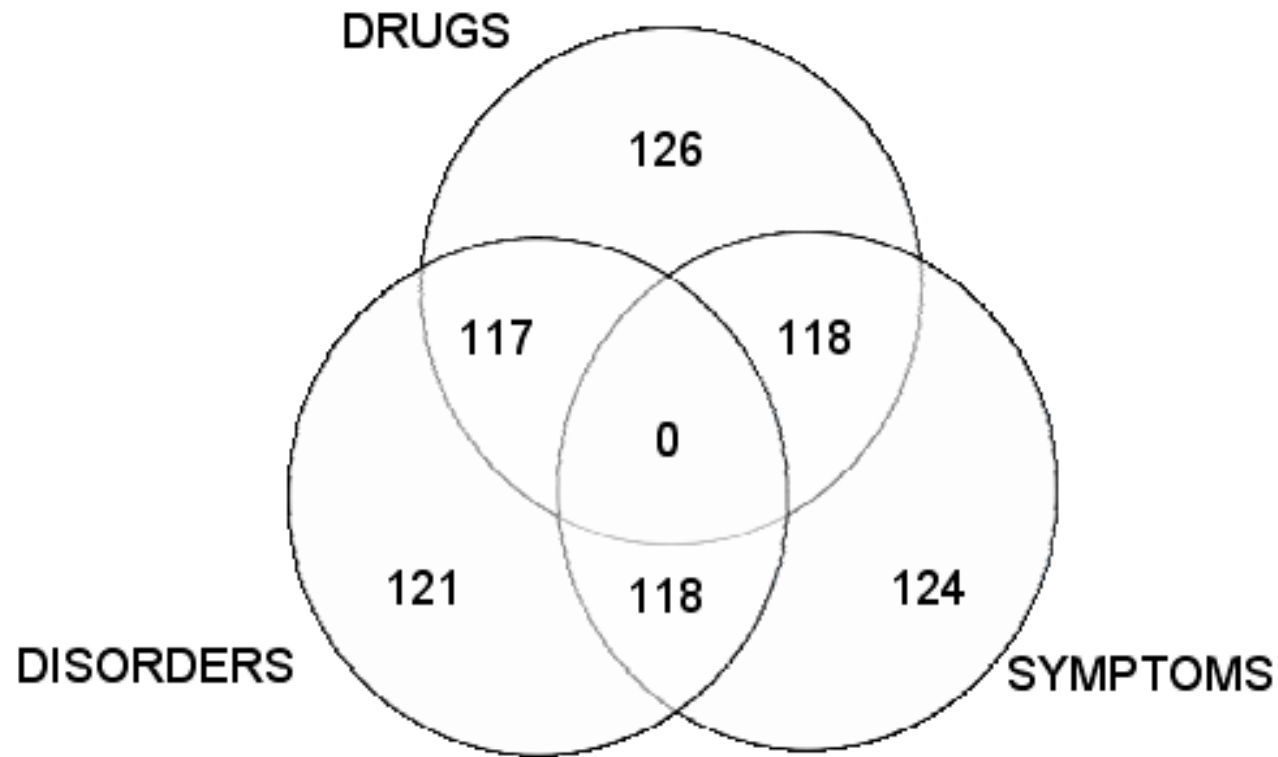
Mini-Mayo-SRS

- 29 pairs with higher agreement
 - mean r (physicians) = 0.68
 - mean r (coders) = 0.78
- This set may be used in
 - system development (e.g., regression testing)
 - and to perform rough comparisons of relative performance of semantic relatedness measures

Available Reference Standards

- UMN SRS
 - result of a psycholinguistic study
 - annotators: 8 medical residents (2 women, 6 men; mean age – 30 y.o.)
 - continuous judgments using a touch screen
 - 4 second time limit
 - single word UMLS concepts from within three semantic types (drugs, diseases and symptoms) and across types
 - total pairs: 724

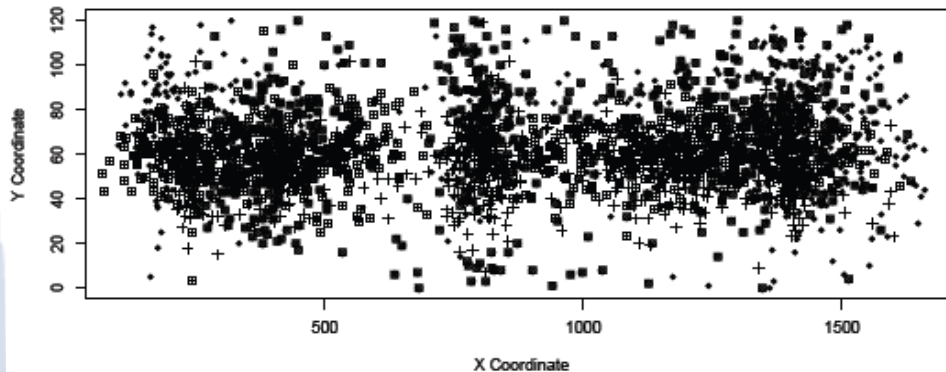
Distribution of term pairs UMN SRS



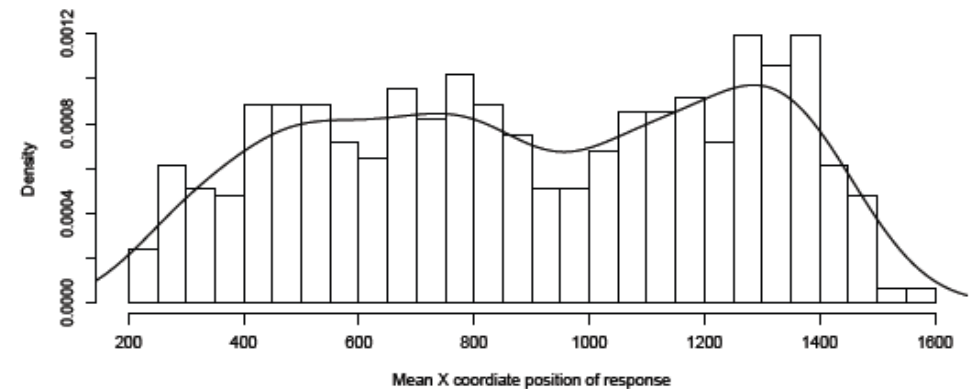
Responses on relatedness task

587 pairs with ICC: 0.5

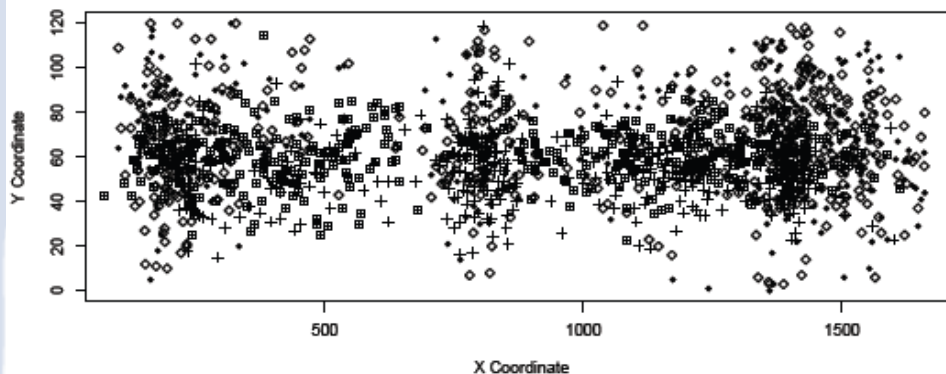
All responses on semantic relatedness task



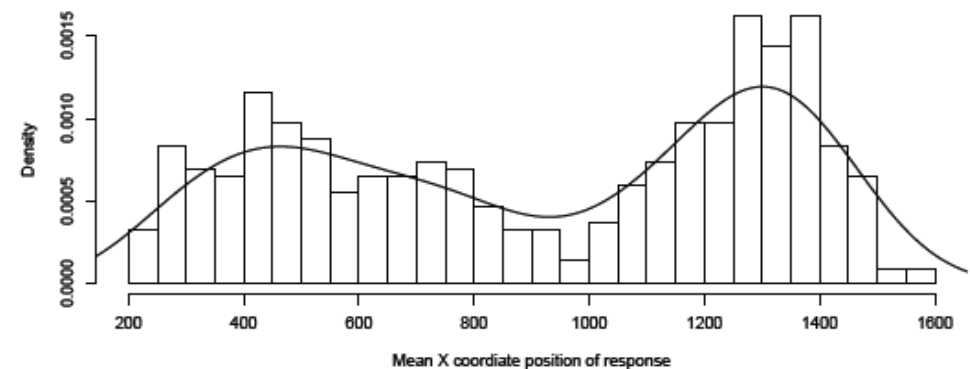
Distribution of mean responses BEFORE reduction



Matched and reduced responses on semantic relatedness task



Distribution of mean responses AFTER reduction

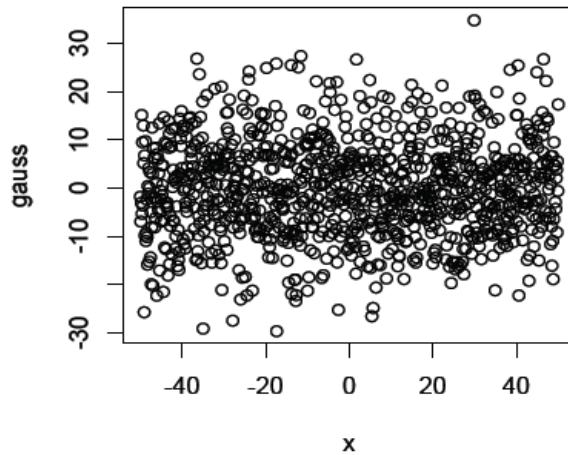


430 pairs with ICC of 0.73

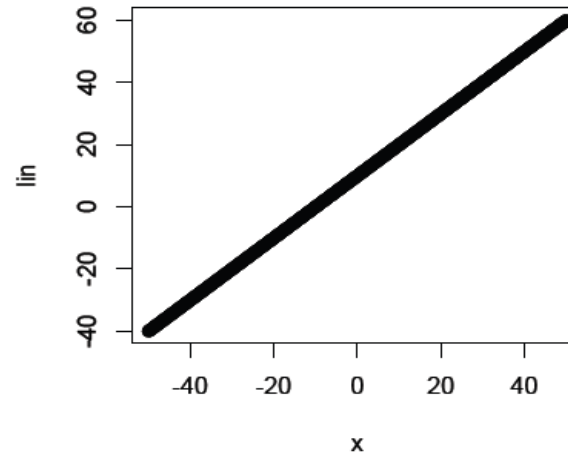
Statistical Considerations

- Pearson's or Spearman's Correlation?
 - depends on the data
- Pearson assumptions
 - values are normally distributed
 - cases are independent
 - relationship between variables is linear
- Spearman assumptions:
 - values are monotonically distributed (important!)

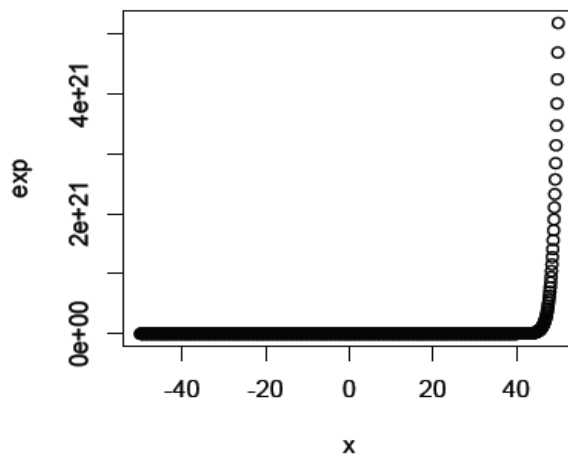
Example



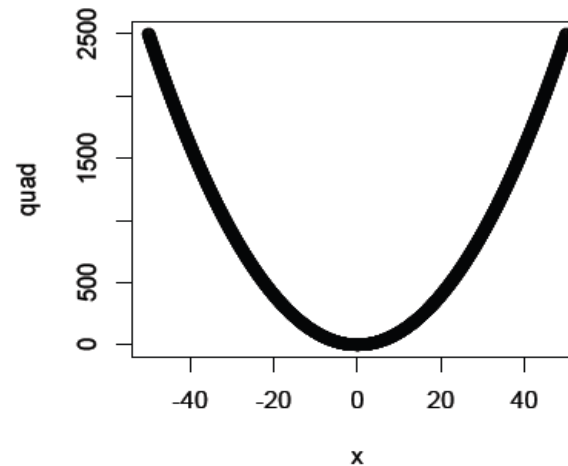
Pearson: 0; Spearman: 0



Pearson: 1; Spearman: 1



Pearson: 0.24; Spearman: 1

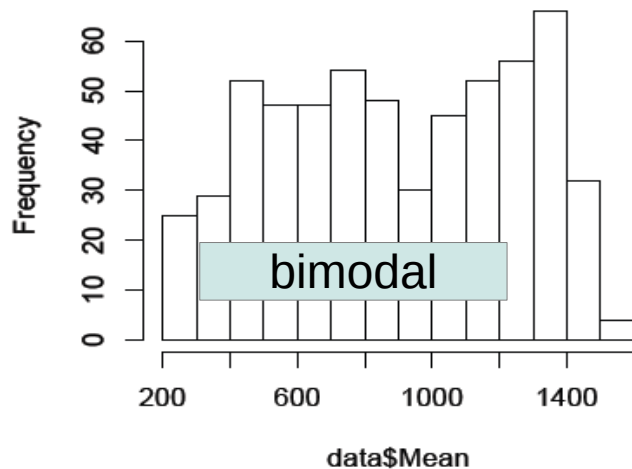


$\text{lm}(\text{quad} \sim \text{poly}(x,2))$
 $R^2 = 1$

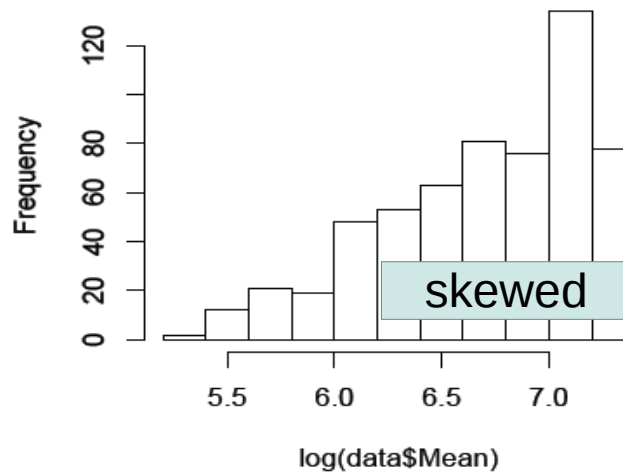
Pearson: 0; Spearman: 0

UMN SRS characteristics

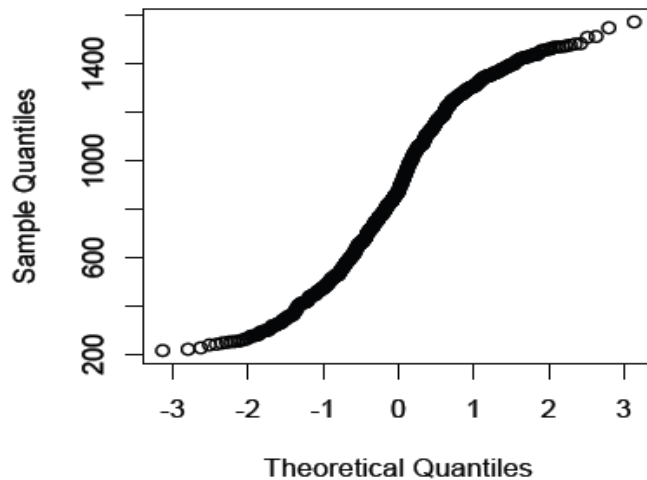
Histogram of data\$Mean



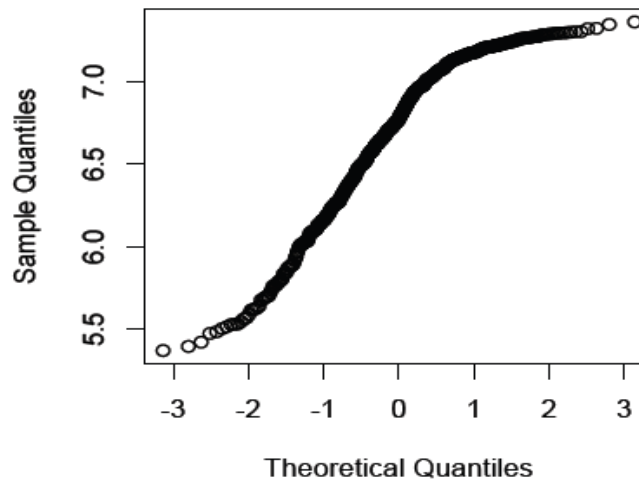
Histogram of log(data\$Mean)



Normal Q-Q Plot



Normal Q-Q Plot



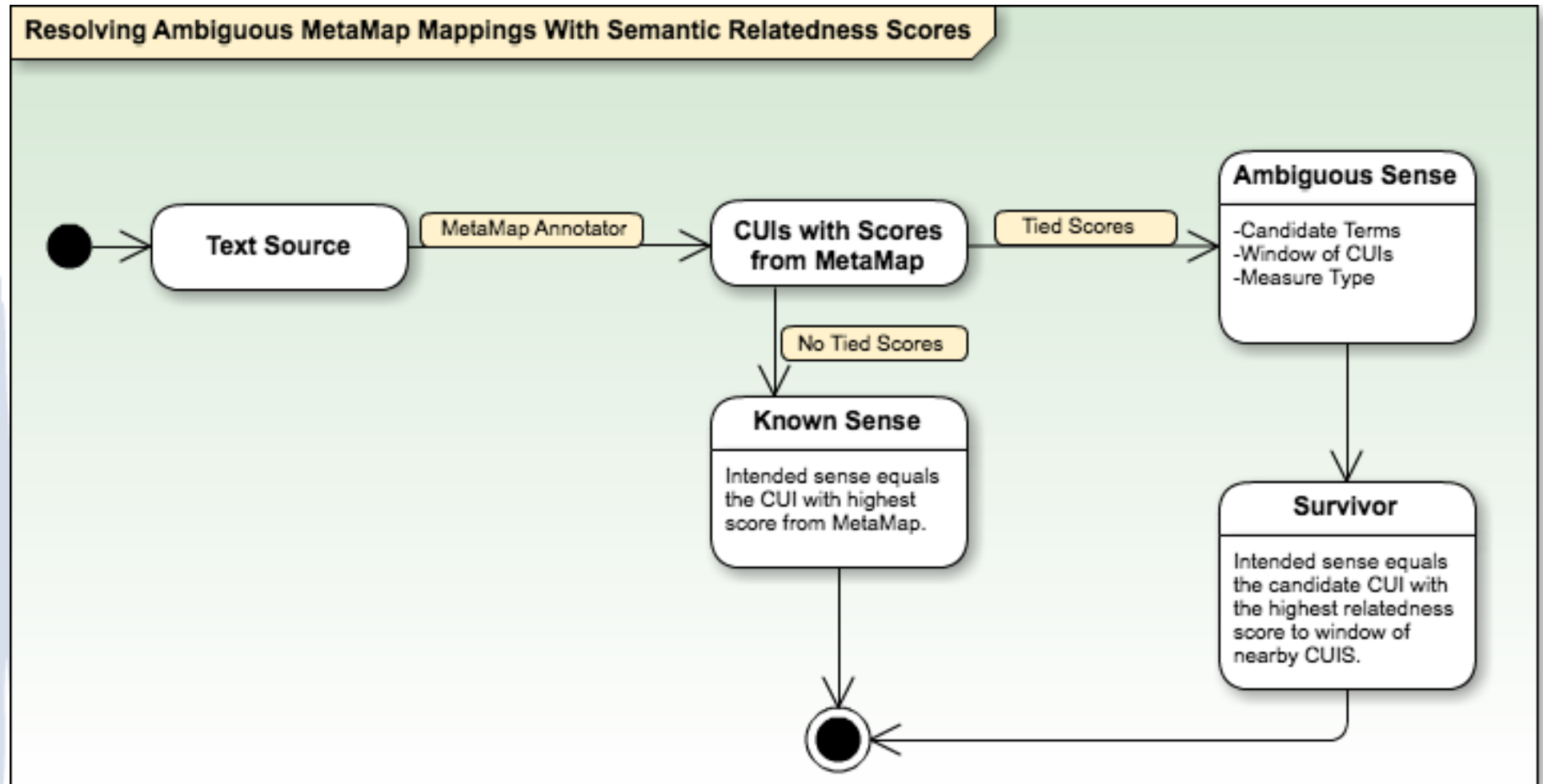
Statistical Considerations

- Direct Evaluation against Mayo or UMN SRS:
 - Spearman rank correlation
 - non-parametric methods
 - consider non-linearity
- Indirect evaluations:
 - precision/recall
 - confidence intervals
 - testing for differences between means

An Example Application

- BioMEDICUS and SenseRelate for acronym disambiguation
 - Biomedical Information Collection and Understanding System (BiomedICUS)
 - open source, UIMA-based
 - <http://code.google.com/p/biomedicus/>
 - SenseRelate
 - broad-coverage word sense tagger
 - uses semantic relatedness
 - <http://www.d.umn.edu/~tpederse/senserelate.html>

Example Application



References

- Pakhomov, S., Pedersen, T., McInnes, B., Melton, G., Ruggieri, A., Chute, C. (2010). Towards a framework for developing semantic relatedness reference standards. *Journal of Biomedical Informatics*. 44(2):251-165.
- Pedersen, T., Pakhomov, S., Patwardhan, S. (2006). Measures of Semantic Similarity and Relatedness in the Biomedical Domain. *Journal of Biomedical Informatics*; 40(3), 288-299.
- Liu, Y., McInnes, B., Pedersen, T., Melton, G.B., Pakhomov, S. (2012). Semantic Relatedness Study Using Second Order Co-Occurrence Vector Computed by Biomedical Corpora, UMLS and WordNet. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI 2012) (January, 2012). Miami, Florida. (in press)
- McInnes, B., Pedersen, T., Liu, Y., Pakhomov, S., Melton, G.B. (2011). Knowledge-based Method for Determining the Meaning of Ambiguous Biomedical Terms Using Information Content Measures of Similarity. In Proceedings of the *American Medical Informatics Symposium (November 2011)*, (in press).
- McInnes, B., Pedersen, T., Liu, Y., Pakhomov, S., Melton, G. (2011). Using Second-order Vectors in a Knowledge-based Method for Acronym Disambiguation. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011) (June 2011). Portland, OR, pp. 145 – 153.
- Pakhomov, S. McInnes, B., Adam, T., Liu, Y., Pedersen, T., Melton, G. (2010) Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. In Proceedings of the *American Medical Informatics Symposium (November 2010)*, pp. 572-576.
- Melton, G., Moon, R. McInnes, B., Pakhomov, S. (2010) Automated Identification of Synonyms in Biomedical Acronym Sense Inventories. In Proceedings of Louhi 02 Workshop at the North American Association of Computational Linguistics, Los Angeles, CA.
- McInnes, B. Pedersen, T. & Pakhomov, S. (2007). Determining the Syntactic Structure of Medical Terms in Clinical Notes. In Proceedings of the BioNLP workshop at the Association for Computational Linguistics Symposium, June 2007, Prague, Czech Republic, pp. 9-16
- Pakhomov, S., Pedersen, T., & Chute, C. G. (2005). Abbreviation and Acronym Disambiguation in Clinical Discourse. In Proceedings of the American Medical Informatics Association Annual Symposium, October 2005, Washington, DC., pp. 589-593
- Rubenstein H, Goodenough J. Contextual correlates of synonymy. *Communications of the ACM* 1965;8:627–33.
- Miller G, Charles W. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 1991;6(1):1–28.
- Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420–8.