Knowledge-based Method to Determine the Meaning of Ambiguous Biomedical Terms Using Measures of Semantic Similarity and Relatedness

Bridget T. McInnes
Department of Computer Science
Virginia Commonwealth University

OBJECTIVE OF THIS WORK

- Develop and evaluate a method than can
  disambiguate terms in biomedical text by
  exploiting similarity and relatedness information
  extrapolated from the Unified Medical Language System

- Evaluate the efficacy of similarity measures
  and relatedness measures
  for Word Sense Disambiguation, WSD

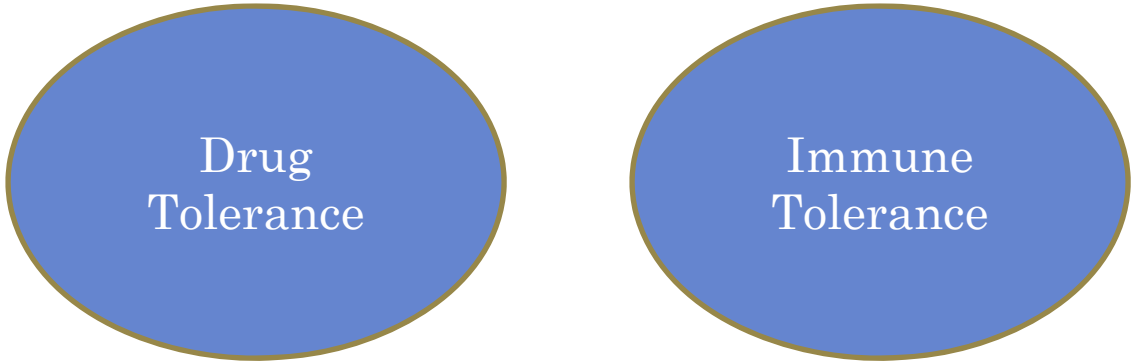
OVERVIEW

- Part I: WSD
- Part II: WSD Algorithm
- Part III: Semantic similarity and relatedness measures
- Part IV: Evaluation Framework
- Part V: Results

WORD SENSE DISAMBIGUATION

Determine the appropriate sense of a term from its context.

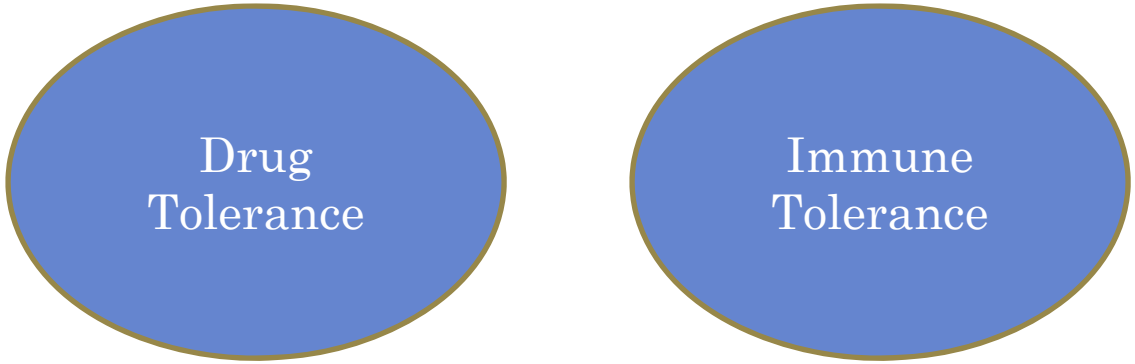
TERM: tolerance



WORD SENSE DISAMBIGUATION

Determine the appropriate sense of a term from its context.

Busprione attenuates tolerance to morphine in mice with skin cancer

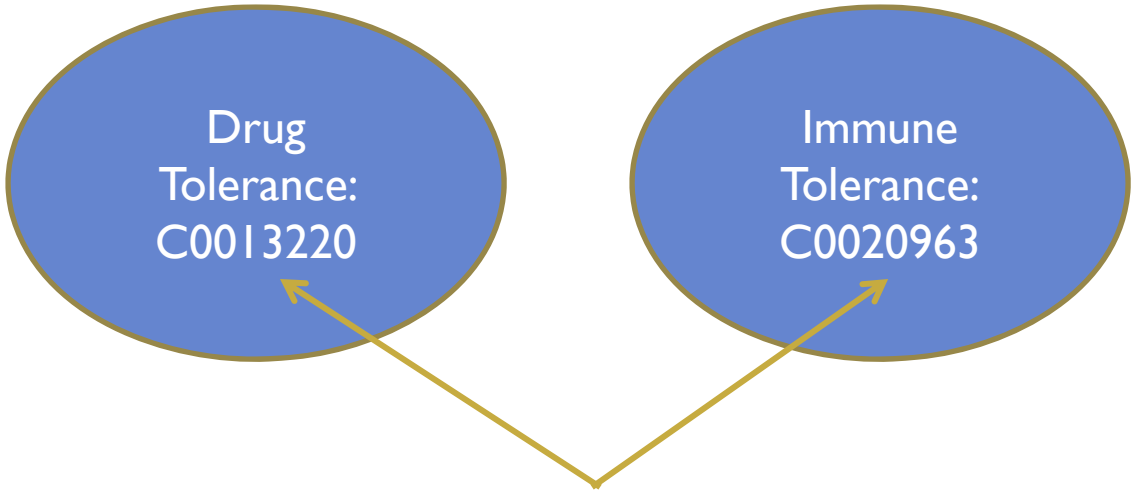


SENSE INVENTORY: UNIFIED MEDICAL LANGUAGE SYSTEM

- Unified Medical Language Sources (UMLS)
  - Semantic Network
  - Metathesaurus
    - ~2 million biomedical and clinical concepts; integrated semi-automatically
    - CUIs (Concept Unique Identifiers), linked:
      - Hierarchical: PAR/CHD and RB/RN
      - Non-hierarchical: SIB, RO
    - Sources viewed together or independently
      - Medical Subject Heading (MSH)
  - SPECIALIST Lexicon
    - Biomedical and clinical terms, including variants

WORD SENSE DISAMBIGUATION

Busprione attenuates tolerance to morphine in mice with skin cancer



PURPOSE

- MetaMap, Aronson 2001
  - Concept mapping system
  - Maps terms to CUIS in the UMLS based on patterns
  - Does not perform WSD

  - Backbone of two other systems:
    - Medical Text Indexer (MTI): CUI recommender for the purpose of indexing biomedical journal articles
    - SemRep: automatically identifies relationships between terms in biomedical text
      - Drug X *treats* Disease Y

WSD ALGORITHM: SENSERELATE

SENSERELATE ALGORITHM

- Each possible sense of a **target word** is assigned a score [sum similarity between it and its surrounding terms]
- Assign target word the sense with highest score

- Proposed by Patwardhan and Pedersen 2003 using WordNet

- UMLS::SenseRelate is a modification of this algorithm using information from the UMLS

NEXT UP: an example

SENSERELATE EXAMPLE

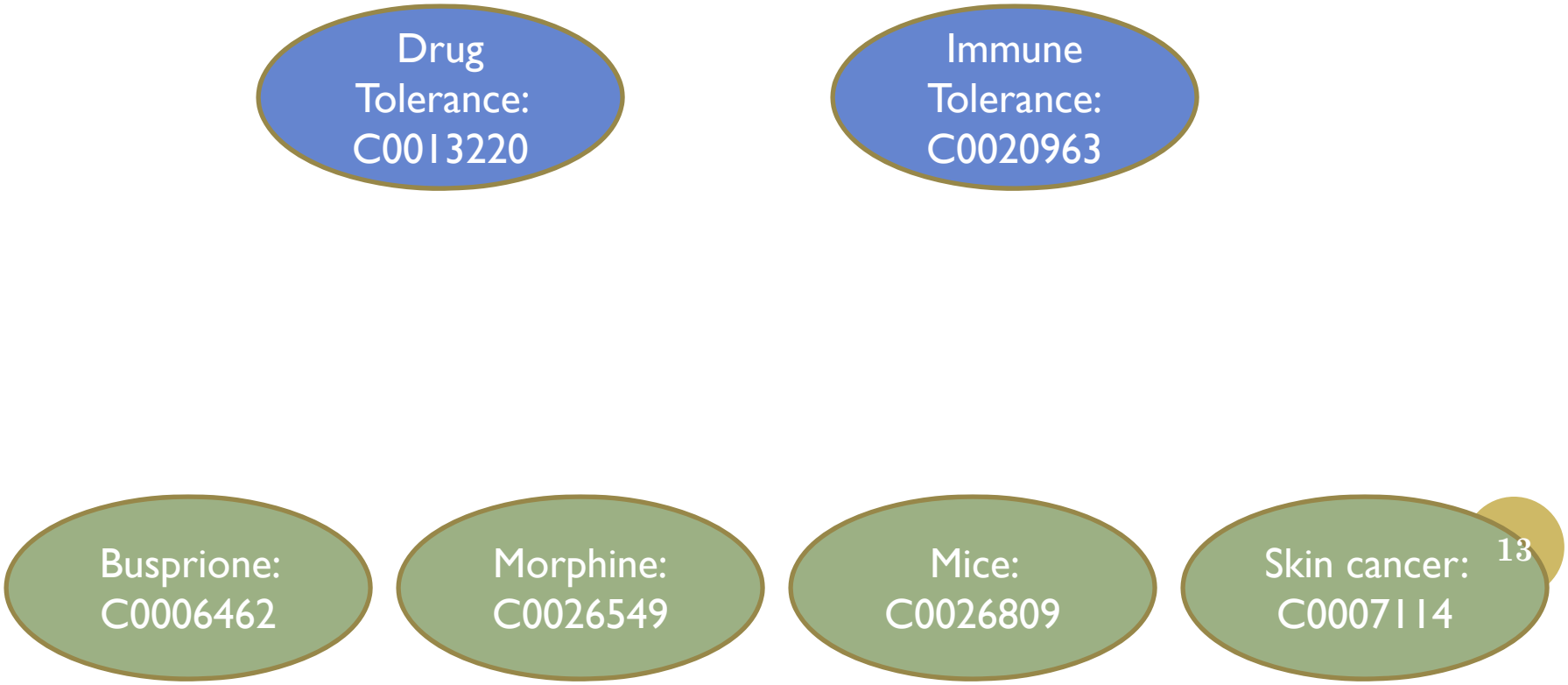
Busprione attenuates tolerance to morphine in mice with skin cancer

SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer

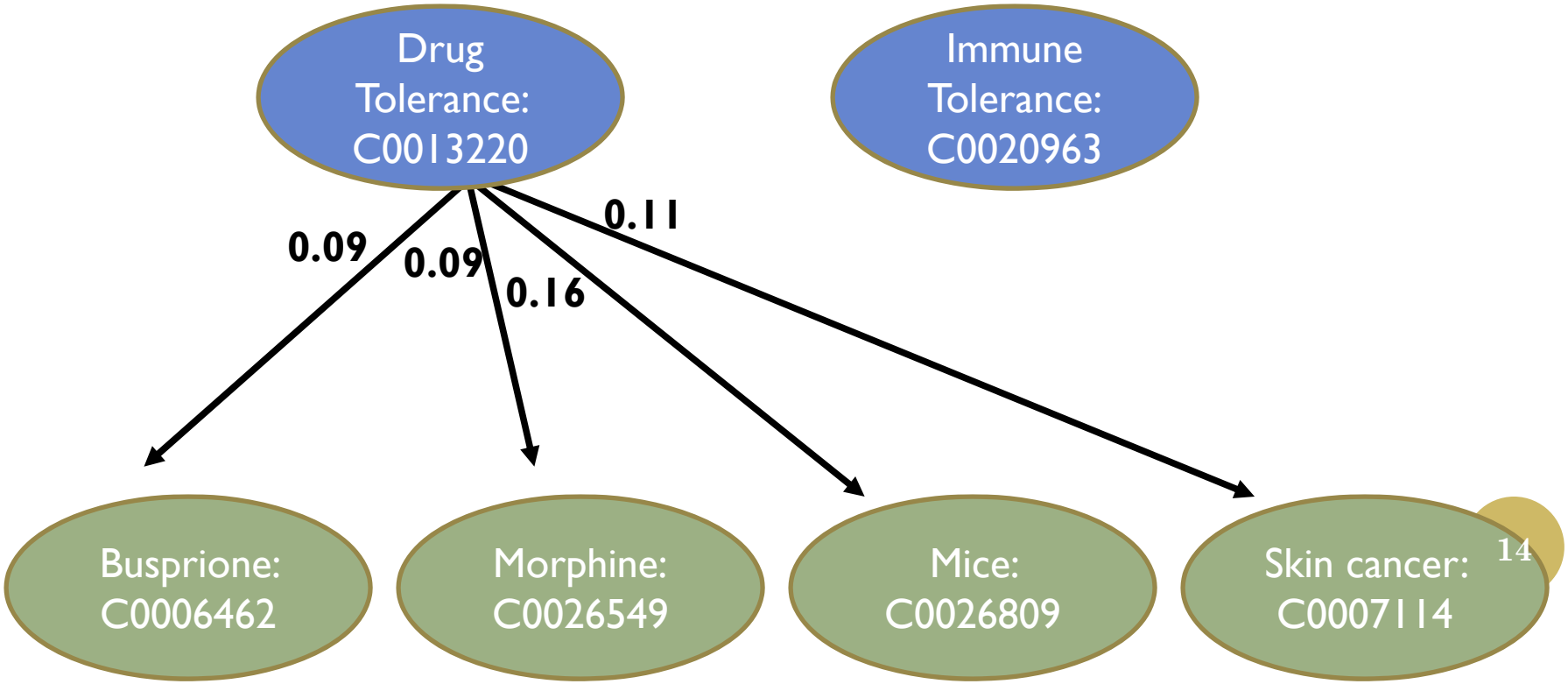
SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer



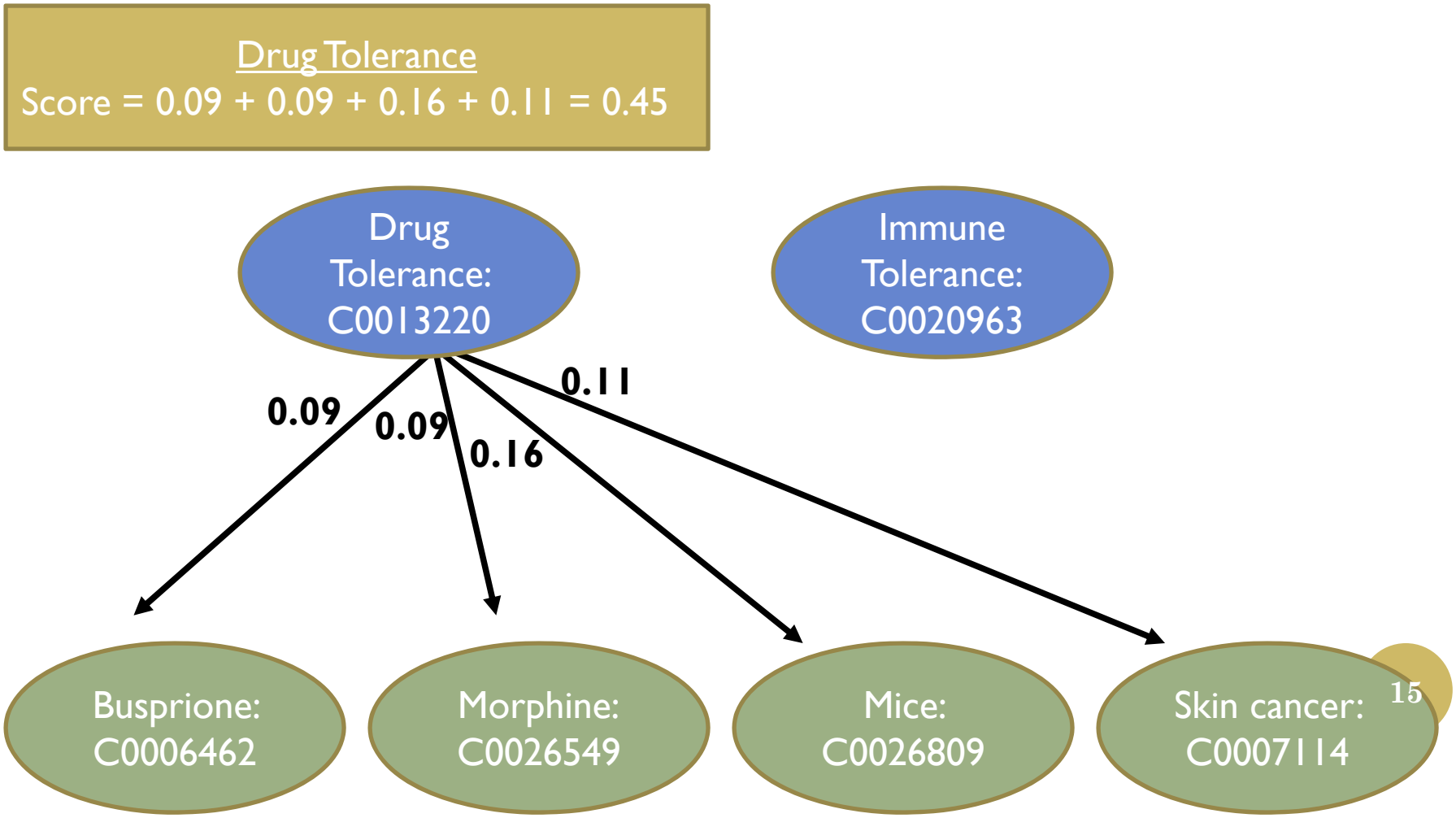
SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer



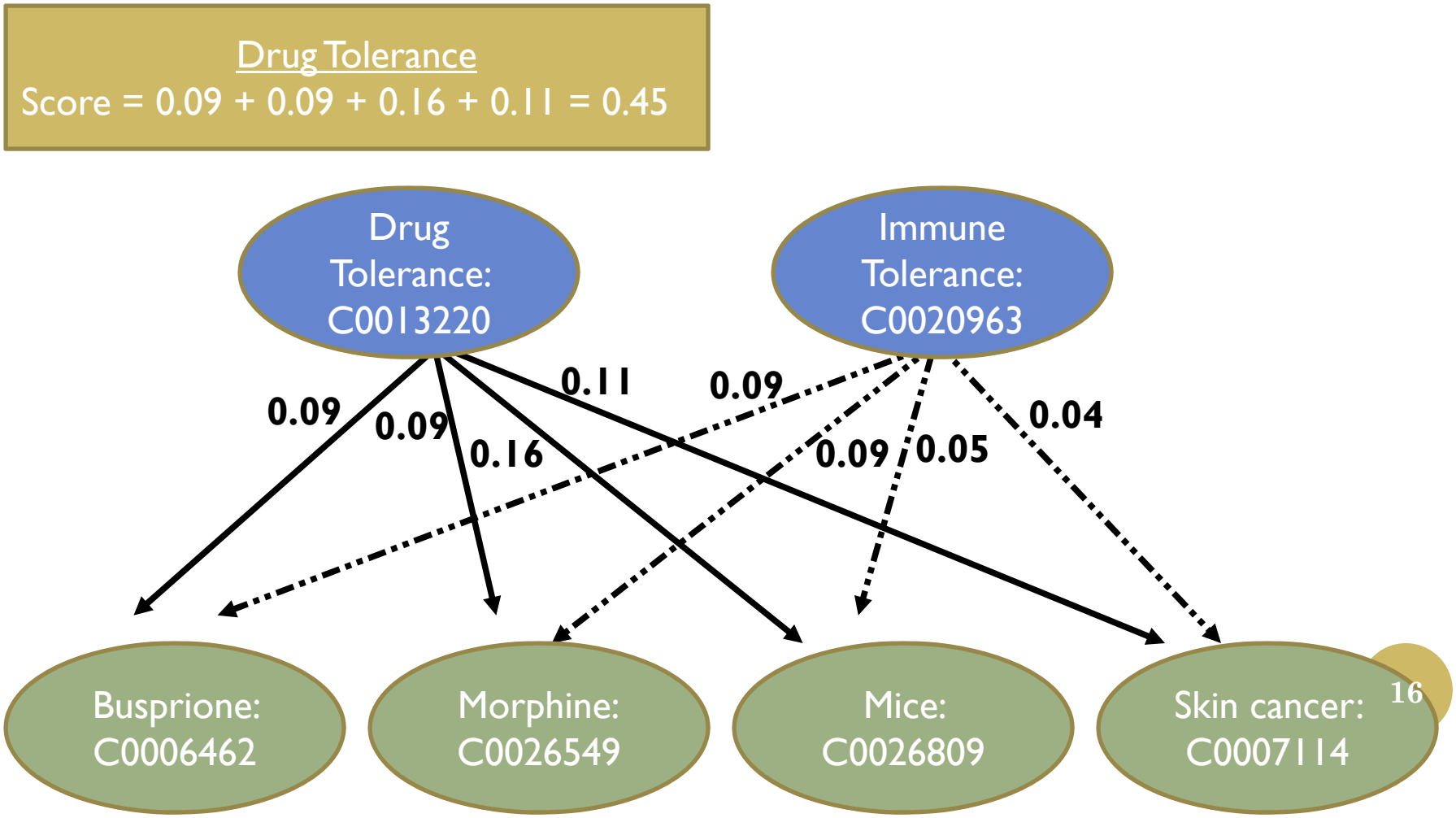
SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer



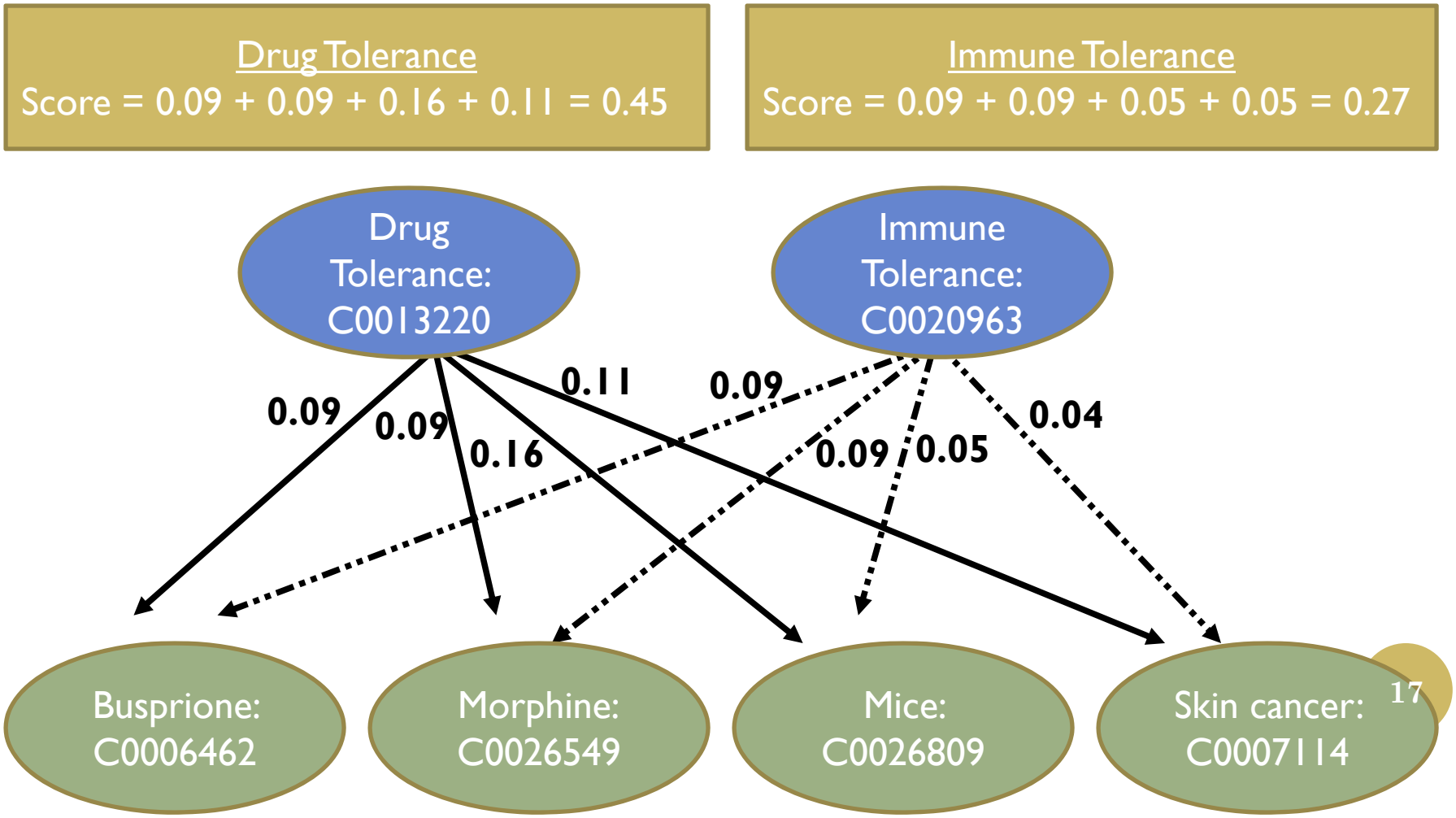
SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer



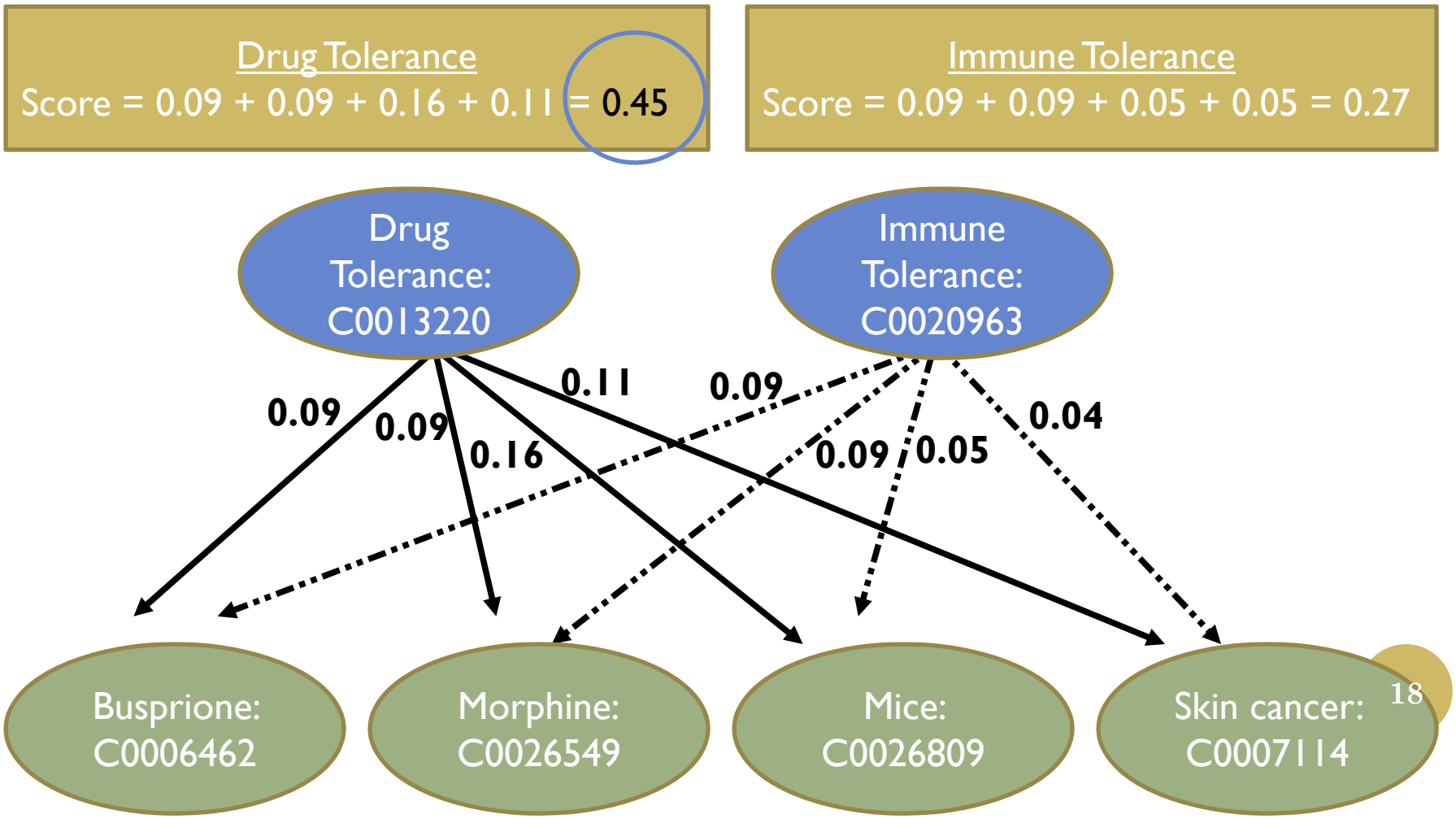
SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer



SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer



SENSE RELATE **ASSUMPTION**

An ambiguous word is often used in the sense that is most similar to the sense of the terms that surround it

IDENTIFYING THE CONCEPTS OF THE SURROUNDING TERMS

Use the SPECIALIST LEXICON to identify the terms and map the terms doing a string match to the MRCONSO table in the UMLS Metathesaurus

IDENTIFYING THE CONCEPTS OF THE SURROUNDING TERMS

Use the SPECIALIST LEXICON to identify the terms and map the terms doing a string match to the MRCONSO table in the UMLS Metathesaurus

Busprione attenuates tolerance to morphine in mice with skin cancer

IDENTIFYING THE CONCEPTS OF THE SURROUNDING TERMS

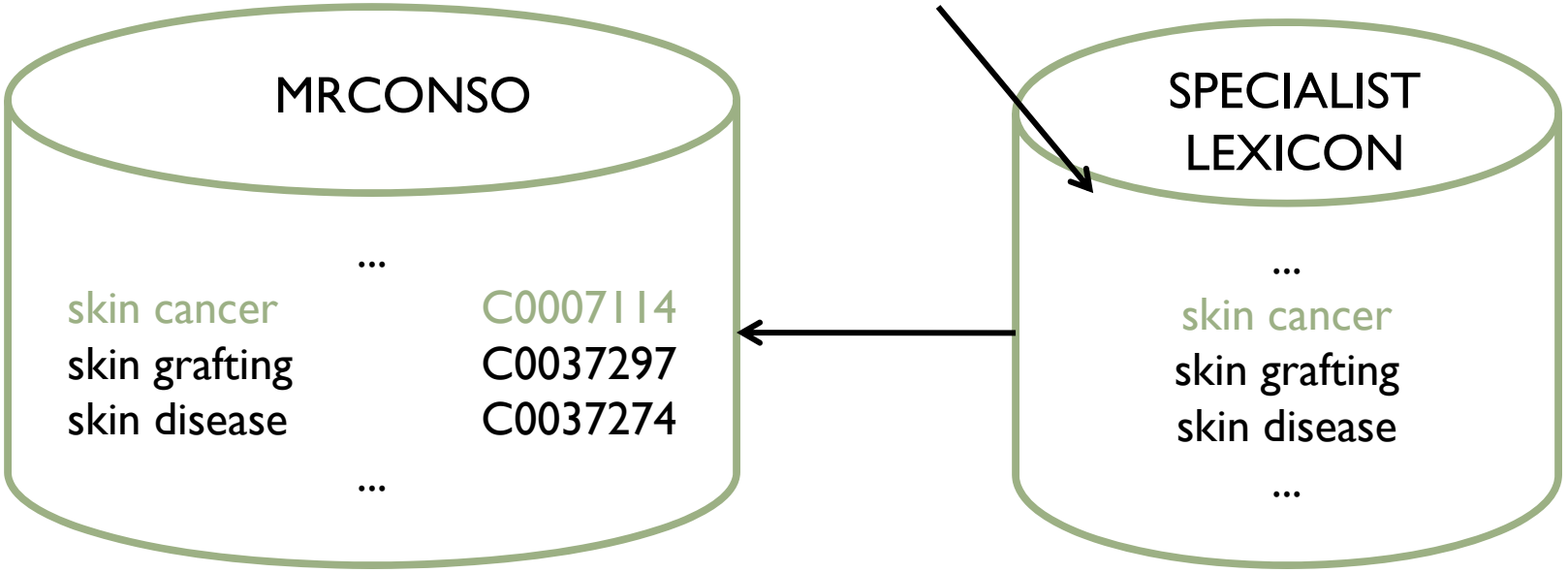
Use the SPECIALIST LEXICON to identify the terms and map the terms doing a string match to the MRCONSO table in the UMLS

Busprione attenuates tolerance to morphine in mice with skin cancer

IDENTIFYING THE CONCEPTS OF THE SURROUNDING TERMS

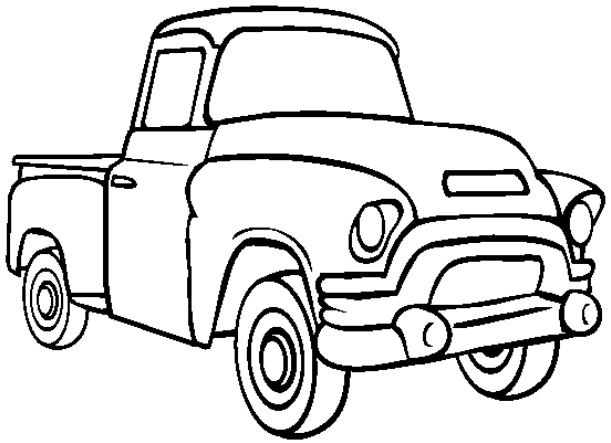
Use the SPECIALIST LEXICON to identify the terms and map the terms doing a string match to the MRCONSO table in the UMLS

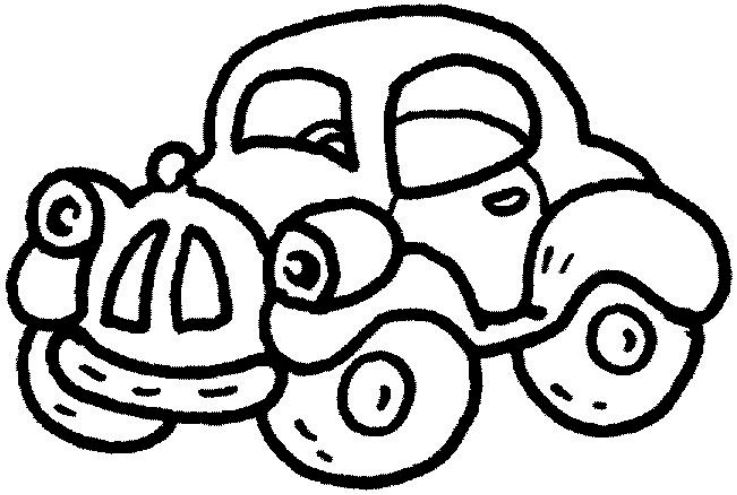
Busprione attenuates tolerance to morphine in mice with skin cancer

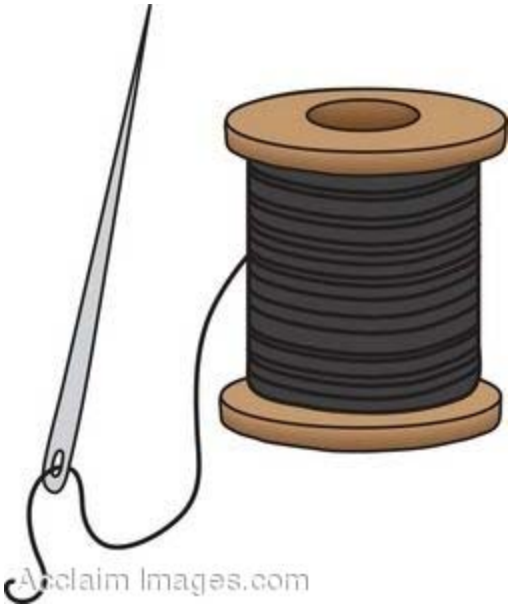


SEMANTIC SIMILARITY AND RELATEDNESS

SEMANTIC SIMILARITY AND RELATEDNESS







SEMANTIC SIMILARITY AND RELATEDNESS MEASURES

- Semantic similarity measures
  - Path-based
  - Information content (IC)-based
- Relatedness measures

PATH-BASED SIMILARITY MEASURES

- Use only the path information obtained from a taxonomy

PATH-BASED SIMILARITY MEASURES

- Use only the path information obtained from a taxonomy

- Path measure
  - sim(c1,c2) = 1 / minpath(c2,c2)
    - where minpath is the shortest path between the two concepts

PATH-BASED SIMILARITY MEASURES

- Use only the path information obtained from a taxonomy
- Path measure
  - sim(c1,c2) = 1/minpath(c2,c2)
    - where minpath is the shortest path between the two concepts
- Wu and Palmer, 1994
  - sim(c1,c2) = (2*depth(LCS(c2,c2))) / (depth(c1)+depth(c2))
    - where LCS is the least common subsumer of the two concepts

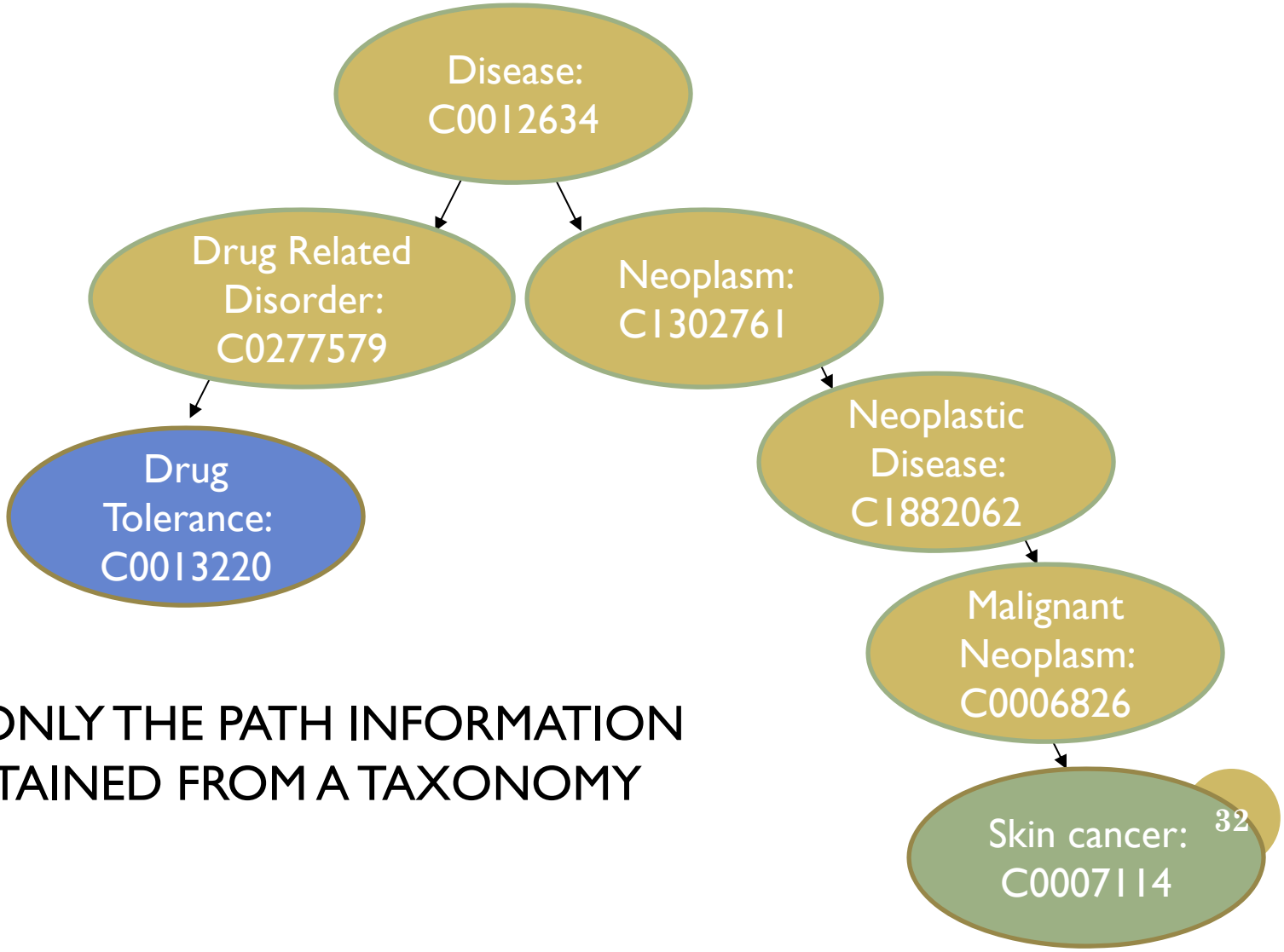
PATH-BASED SIMILARITY MEASURES

- Use only the path information obtained from a taxonomy
- Path measure
  - sim(c1,c2) = 1/ minpath(c2,c2)
    - where minpath is the shortest path between the two concepts
- Wu and Palmer, 1994
  - sim(c1,c2) = (2*depth(LCS(c2,c2))) / (depth(c1)+depth(c2))
    - where LCS is the least common subsumer of the two concepts
- Leacock and Chodorow, 1998
  - sim(c1,c2) = -log(minpath(c1,c2) / (2D))
    - where D is the total depth of the taxonomy

PATH-BASED SIMILARITY MEASURES

- Use only the path information obtained from a taxonomy
- Path measure
  - sim(c1,c2) = 1/ minpath(c2,c2)
    - where minpath is the shortest path between the two concepts
- Leacock and Chodorow, 1998
  - sim(c1,c2) = -log(minpath(c1,c2) / (2D))
    - where D is the total depth of the taxonomy
- Wu and Palmer, 1994
  - sim(c1,c2) = (2*depth(LCS(c2,c2))) / (depth(c1)+depth(c2))
    - where LCS is the least common subsumer of the two concepts
- Nyguen and Al-Mubaid, 2006
  - sim(c1,c2) = log ((2 + minpath(c1,c2) - 1) * (D - depth(LCS(c1,c2))))

PATH-BASED SIMILARITY MEASURES



USE ONLY THE PATH INFORMATION OBTAINED FROM A TAXONOMY

Information content-based Measures

- Incorporate the probability of the concepts
  - IC = -log(P(concept))

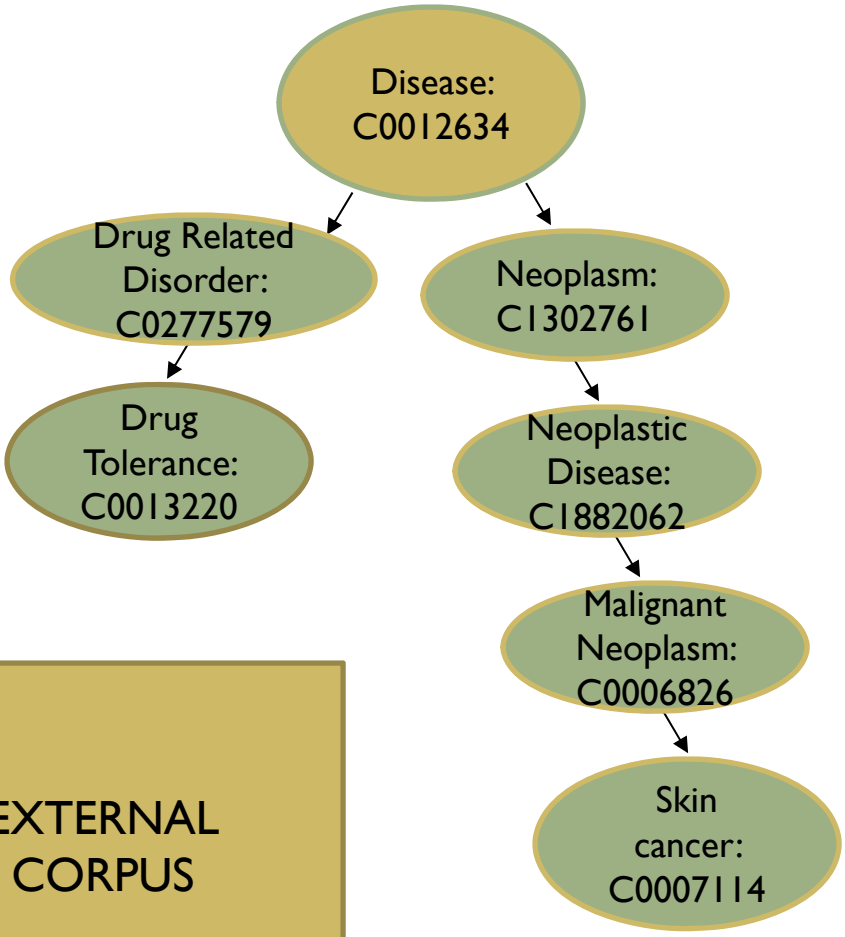
INFORMATION CONTENT-BASED MEASURES

- Incorporate the probability of the concepts
  - IC = -log(P(concept))
- P(concept)
  - Calculated by summing the probability of the concept and the probability of its descendants
  - Probabilities are obtained from an external corpus

PROBABILITY EXAMPLE

P(Disease [C0012634]) =

P(C0012634) +
P(C1302761) +
P(C1882062) +
P(C0006826) +
P(C0007114) +
P(C0277579) +
P(C0013220) ;



EXTERNAL CORPUS

P(C) = freq(C) / N

INFORMATION CONTENT-BASED MEASURES

- Incorporate the probability of the concepts
  - IC = -log(P(concept)
- Resnik, 1995
  - sim(c1,c2) = IC(LCS(c1,c2))

INFORMATION CONTENT-BASED MEASURES

- Incorporate the probability of the concepts
  - IC = -log(P(concept)
- Resnik, 1995
  - sim(c1,c2) = IC(LCS(c2,c2))
- Jiang and Conrath, 1997
  - sim(c1,c2) = 1 / (IC(c1)+IC(c2) – 2* IC(LCS(c1,c2))

INFORMATION CONTENT-BASED MEASURES

- Incorporate the probability of the concepts
  - IC = -log(P(concept)
- Resnik, 1995
  - sim(c1,c2) = IC(LCS(c2,c2))
- Jiang and Conrath, 1997
  - sim(c1,c2) = 1 ÷ (IC(c1)+IC(c2) – 2* IC(LCS(c1,c2))
- Lin, 1998
  - sim(c1,c2) = (2*IC(LCS(c2,c2))) / (IC(c1)+IC(c2))

SIDE NOTE: COMPARISON BETWEEN LIN AND WU & PALMER

- IC-based measure: Lin, 1998
  - sim(c1,c2) = (2*IC(LCS(c2,c2))) / (IC(c1)+IC(c2))

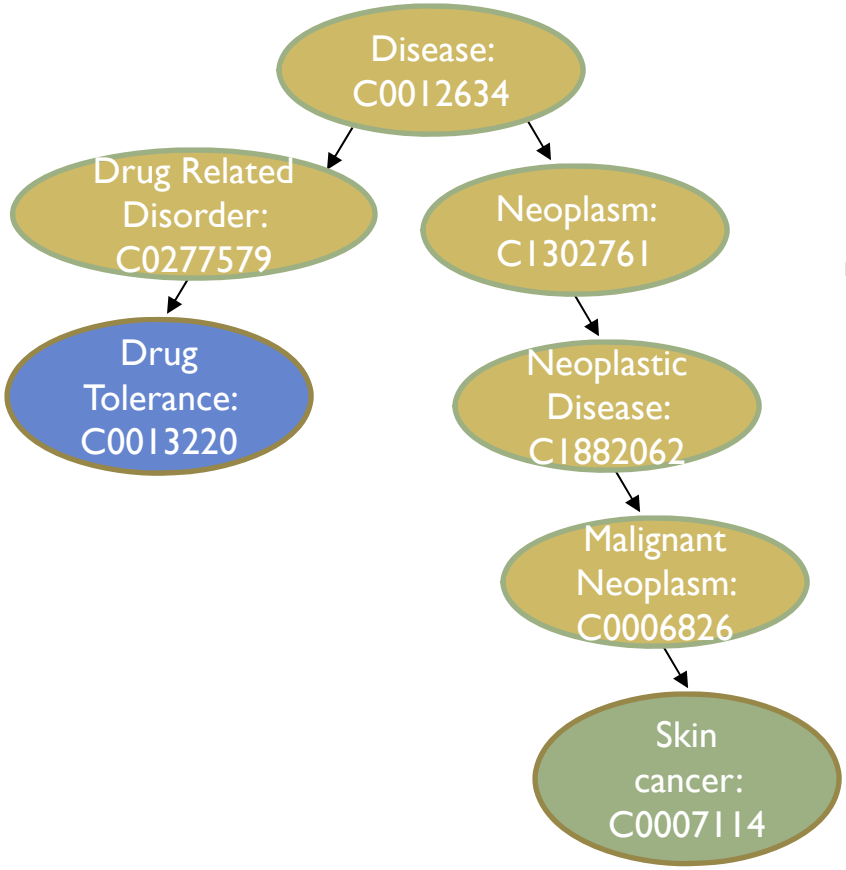
- Path-based measure: Wu and Palmer, 1994
  - sim(c1,c2) = (2*depth(LCS(c2,c2))) / (depth(c1)+depth(c2))

SIDE NOTE: COMPARISON BETWEEN LIN AND WU & PALMER

- IC-based measure: Lin, 1998
  - sim(c1,c2) = (2*IC(LCS(c2,c2))) / (IC(c1)+IC(c2))
- Path-based measure: Wu and Palmer, 1994
  - sim(c1,c2) = (2*depth(LCS(c2,c2))) / (depth(c1)+depth(c2))

IC-BASED SIMILARITY MEASURES

PATH INFORMATION



PROBABILITY OF CONCEPTS

EXTERNAL CORPUS

Relatedness Measures

- Use contextual information describing the concepts

Relatedness Measures

- Use contextual information describing the concepts
  - Lesk (1986)
  - Vector measure
    - Patwardhan and Pedersen (2006)

LESK MEASURE

- Lesk measure: Lesk, 1986
  - rel(c1,c2) = $\sum_{o \in overlap} length(o)^2$
    - where *length* = # words in the term
  - Contextual information representing its term
    - UMLS Definition

LESK MEASURE

- Lesk measure: Lesk, 1986
  - rel(c1,c2) = $\sum_{o \in overlap} length(o)^2$
    - where *length* = # words in the term

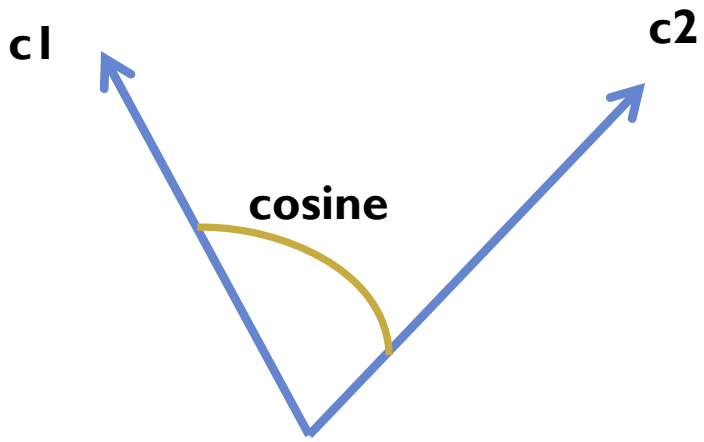
Any of the terminal digits of the hand

One of the terminal digits of the foot

sim(*finger*, *toe*) = $2^2 = 4$

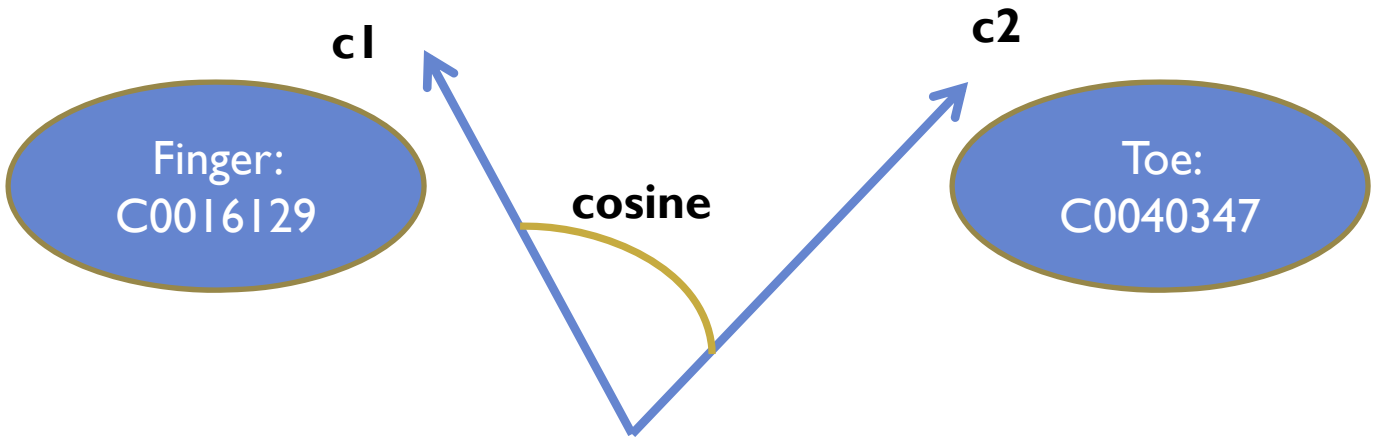
VECTOR MEASURE

- Vector measure: Patwardhan and Pedersen, 2006
  - rel(c1,c2) = cosine(2nd-order vector(c1), 2nd-order vector(c2))



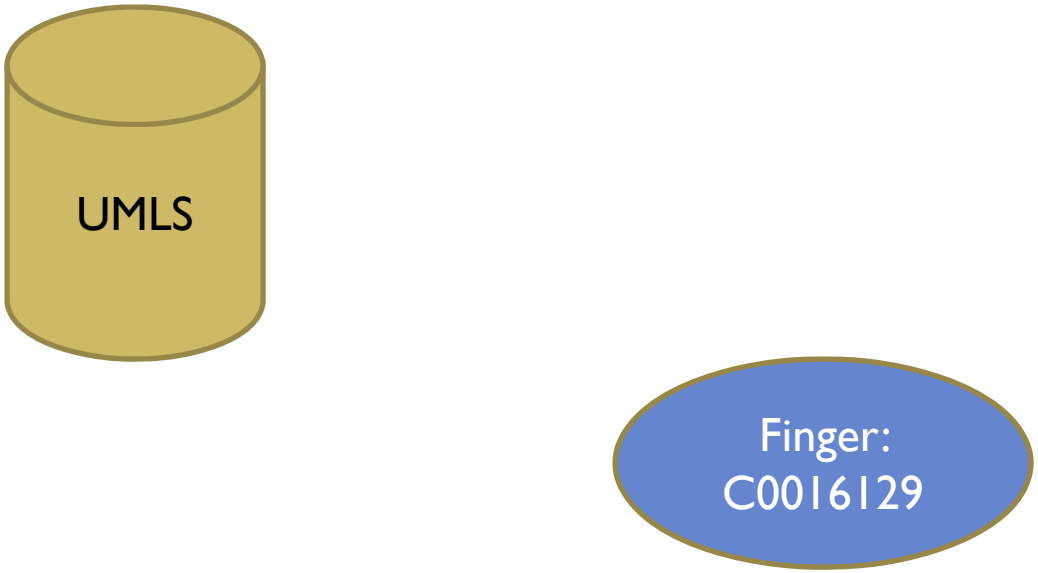
VECTOR MEASURE

- Vector measure: Patwardhan and Pedersen, 2006
  - rel(c1,c2) = cosine(2nd-order vector(c1), 2nd-order vector(c2))

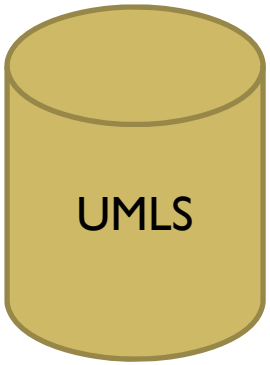


VECTOR MEASURE ALGORITHM

VECTOR MEASURE ALGORITHM

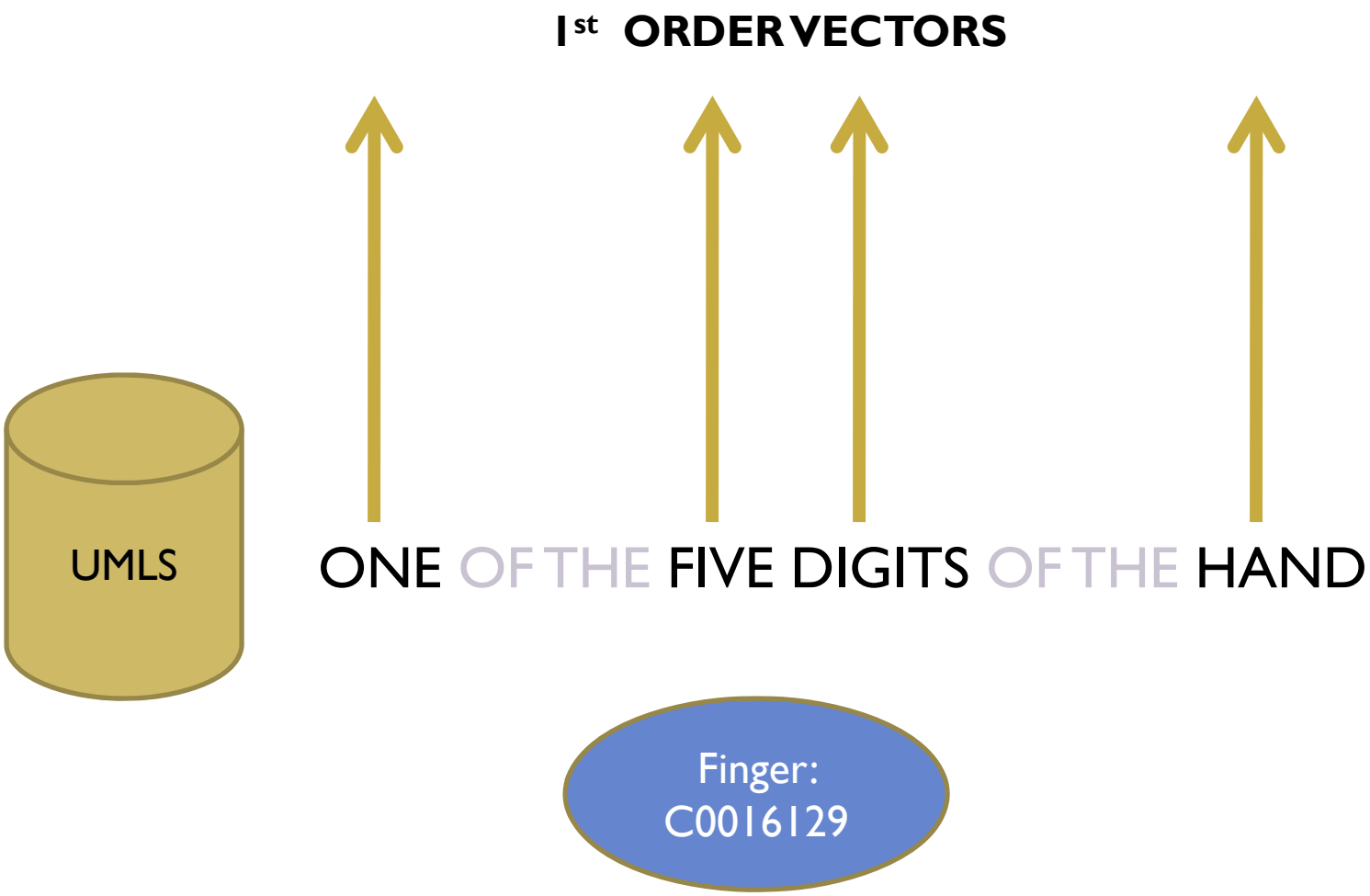


VECTOR MEASURE ALGORITHM

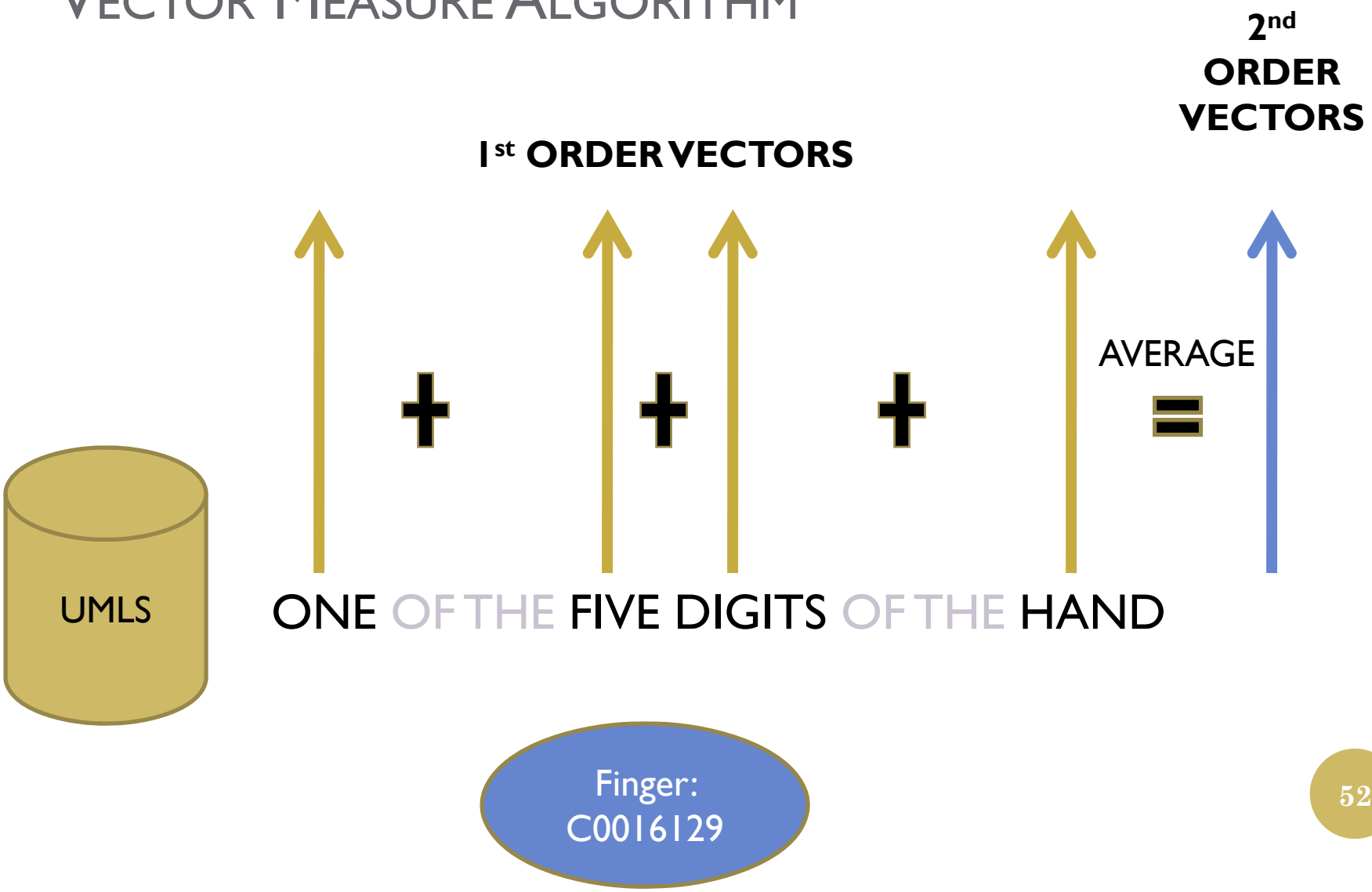


ONE OF THE FIVE DIGITS OF THE HAND

VECTOR MEASURE ALGORITHM



VECTOR MEASURE ALGORITHM



VECTOR MEASURE ALGORITHM

2nd ORDER VECTORS

UMLS

ONE OF THE FIVE DIGITS OF THE HAND

Finger:
C0016129

1ST ORDER VECTORS

EXTERNAL CORPUS

ONE OF THE FIVE DIGITS OF THE HAND

Finger:
C0016129

1ST ORDER VECTORS

EXTERNAL CORPUS

Word 1

Word 2

Word 3

Word 4

Word 5

Word 6

…

Word N

ONE OF THE FIVE DIGITS OF THE HAND

Finger: C0016129

1ST ORDER VECTORS

EXTERNAL CORPUS

	ONE	FIVE	DIGITS	HAND
Word 1	0	2	0	4
Word 2	3	5	0	6
Word 3	4	0	0	0
Word 4	6	0	2	2
Word 5	8	0	0	7
Word 6	10	2	1	0
…	…	…	…	…
Word N	11	1	1	0

Finger: C0016129

1ST ORDER VECTORS

EXTERNAL CORPUS

	ONE	FIVE	DIGITS	HAND
Word 1	0	2	0	4
Word 2	3	5	0	6
Word 3	4	0	0	0
Word 4	6	0	2	2
Word 5	8	0	0	7
Word 6	10	2	1	0
…	…	…	…	…
Word N	11	1	1	0

Finger: C0016129

2ND ORDER VECTOR

(0 + 2 + 0 + 4) / 4
(3 + 5 + 0 + 6) / 4
(4 + 0 + 0 + 0) / 4
(6 + 0 + 2 + 2) / 4
(8 + 0 + 0 + 7) / 4
(10 + 2 + 1 + 0) / 4
…
(11 + 1 + 1 + 0) / 4

ONE FIVE DIGITS HAND

Finger:
C0016129

2ND ORDER VECTOR

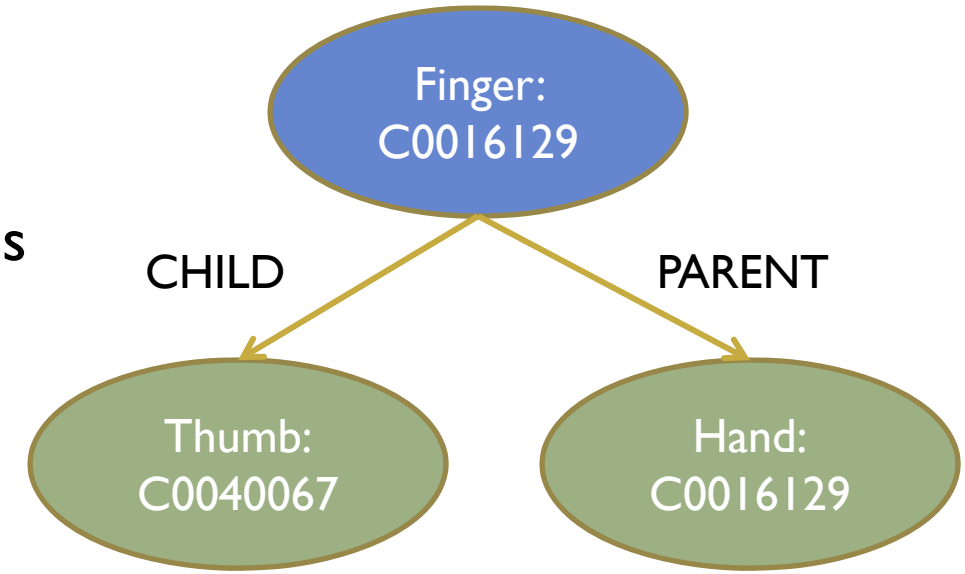
1.5
3.5
1
2.5
3.75
3.25
…
3.25

ONE FIVE DIGITS HAND

Finger:
C0016129

DEFINITIONS FOR RELATEDNESS MEASURE

- UMLS
  - Not all Concepts (CUIs) have a definition
  - Incorporate the definition of its related concepts
    - Parent/Child
    - Narrower/Broader
- Extended Definitions



EXPERIMENTAL FRAMEWORK

EXPERIMENTAL FRAMEWORK

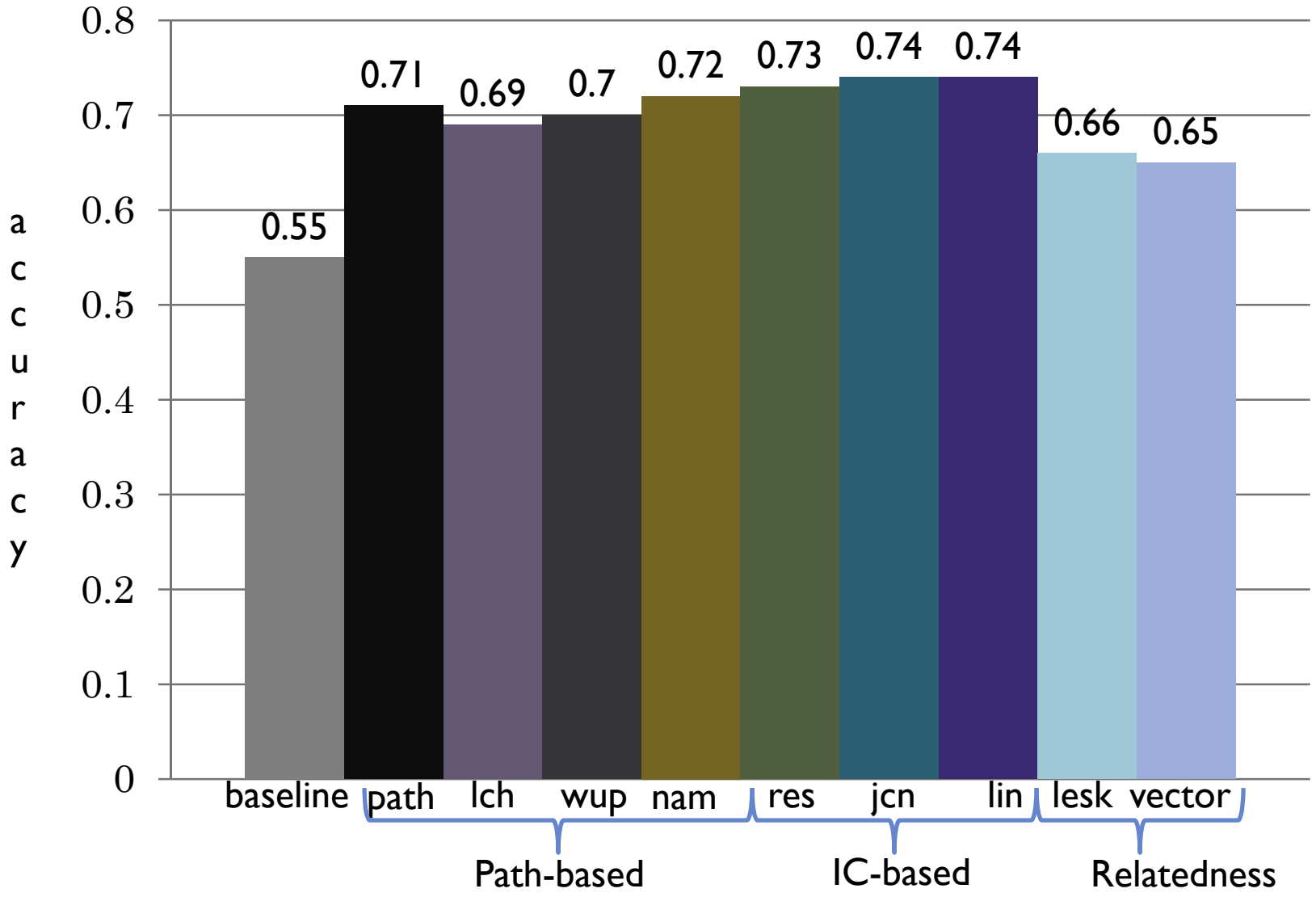
- Use open-source UMLS::Similarity package to obtain the similarity and relatedness between the terms and possible senses in the SenseRelate algorithm
- Path information: parent/child relations in MSH source
- Information content: calculated using the UMLSonMedline dataset created by NLM
  - Consists of concepts from 2009AB UMLS and the frequency they occurred in Medline using the Essie Search Engine (Ide et al 2007)
  - Medline: database of citations of biomedical/clinical articles
- Relatedness information: parent/child and narrower/broader relations from the **entire** UMLS to create the extended definitions

EVALUATION DATA: MSH WSD

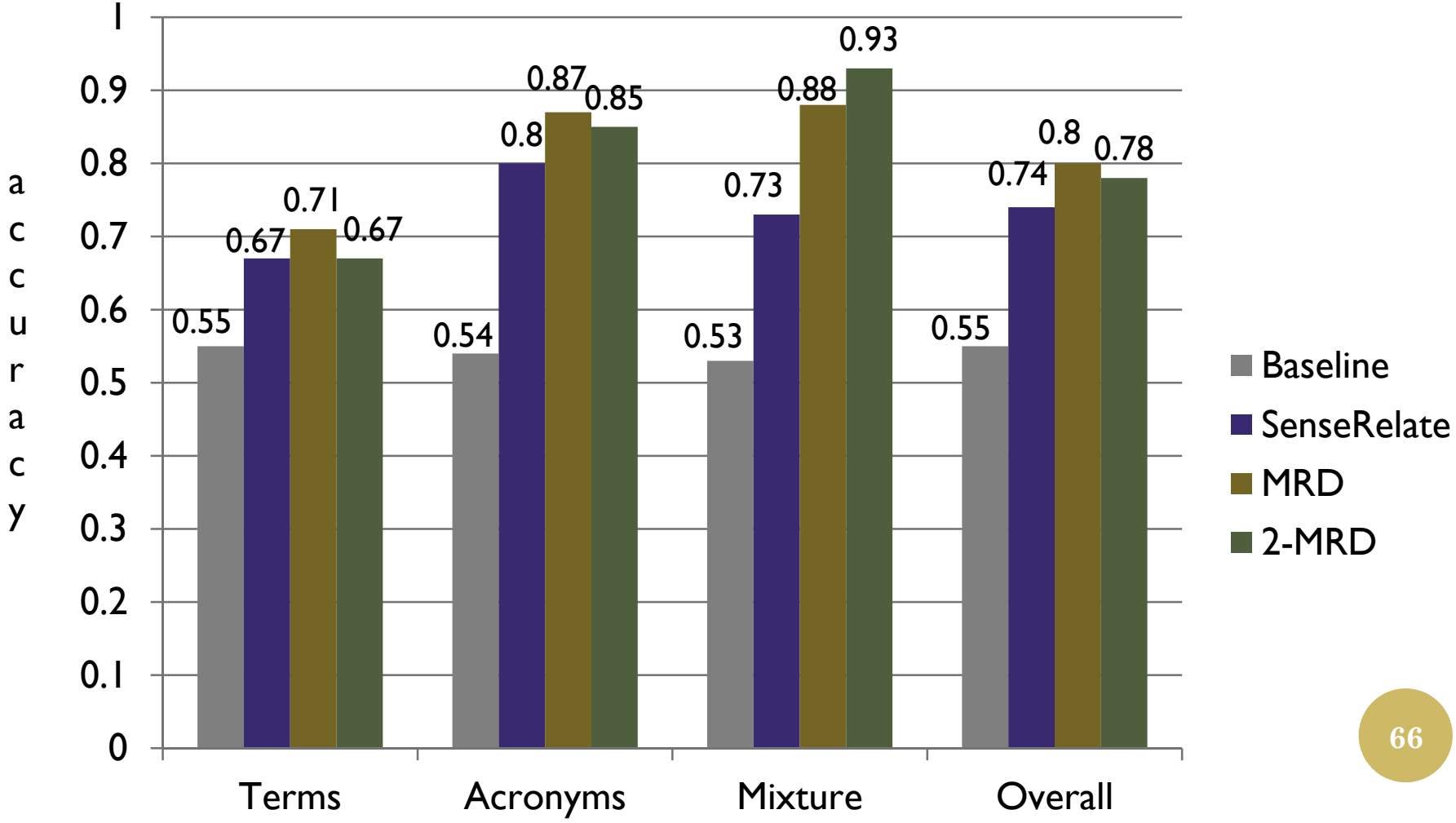
- MSH-WSD dataset (Jimeno-Yepes, et al 2011)
  - 203 target words (ambiguous word) from Medline
    - 106 terms e.g. tolerance
    - 88 acronyms e.g. CA (calcium, california)
    - 9 mixtures e.g. bat (brown adipose tissue)
  - Each target word contains ~187 instances (Medline abstracts)
    - abstract = ~ 500 words
  - Each target word in the instances assigned a concept from MSH by exploiting the manually assigned MSH concepts assigned to the abstract
  - Average of 2.08 possible senses per target word
  - Majority sense over all the target words is 54.5%

RESULTS

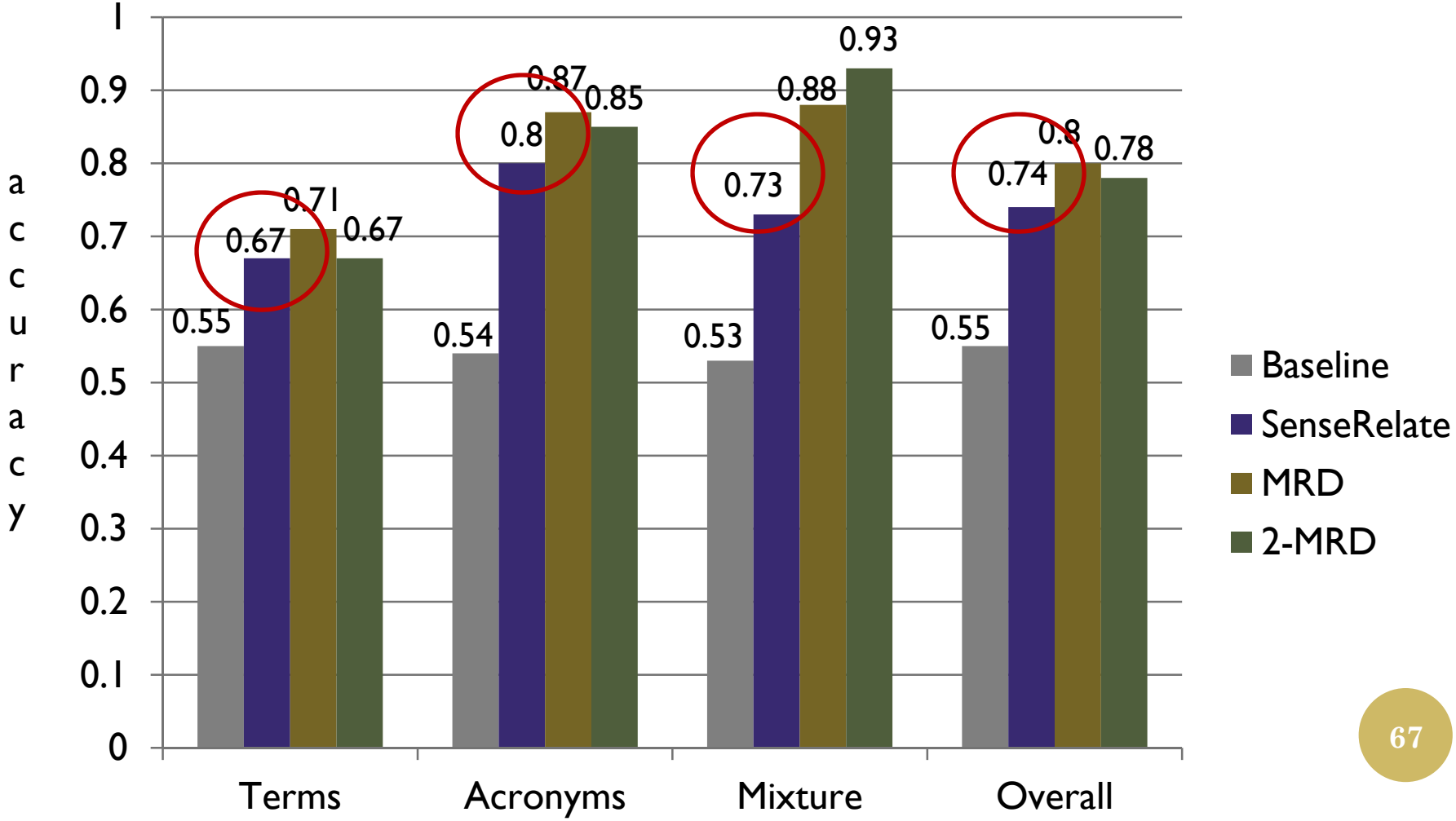
RESULTS OVER MSH-WSD DATASET



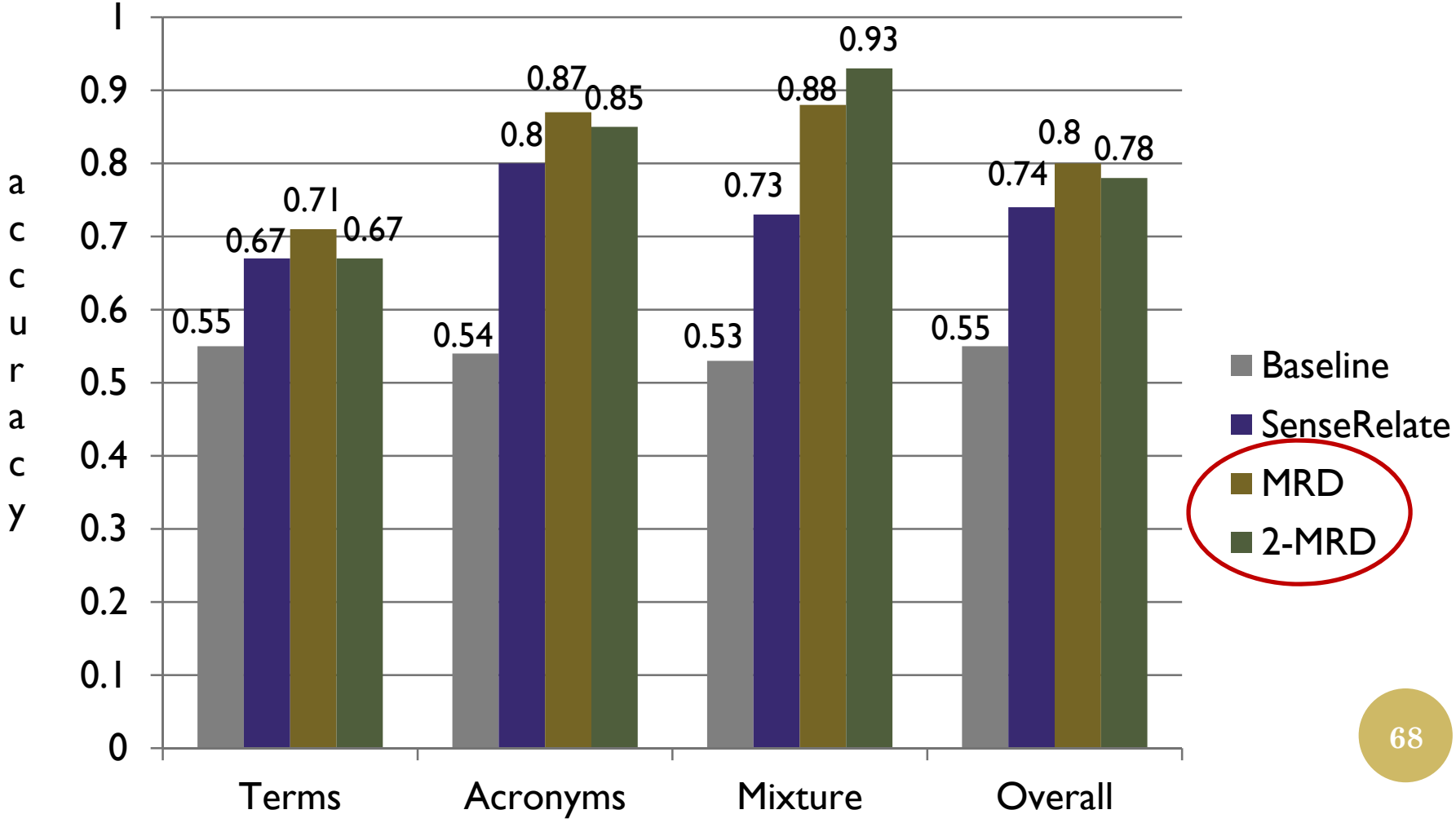
COMPARISON ACROSS SUBSETS OF MSH-WSD



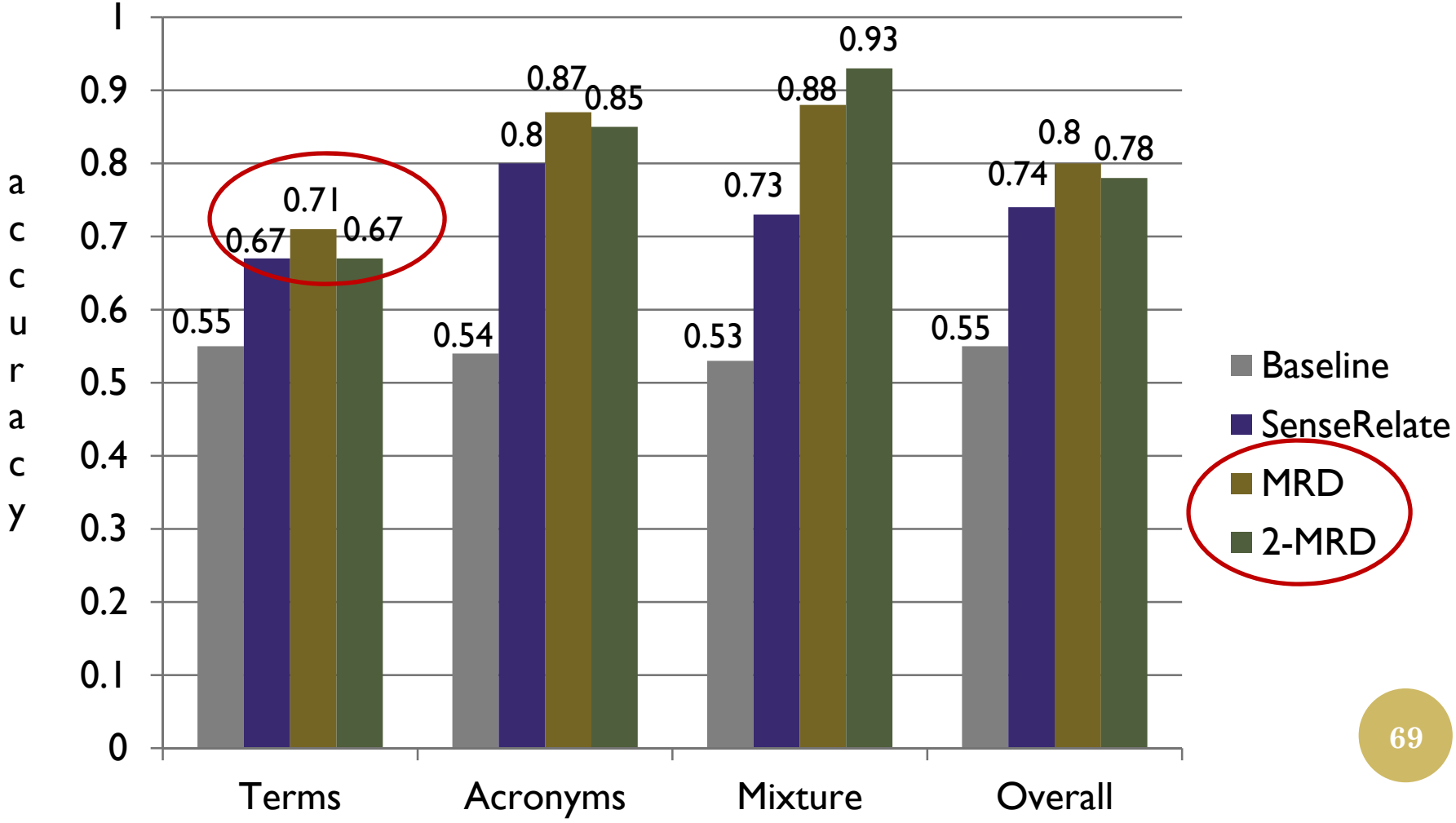
COMPARISON ACROSS SUBSETS OF MSH-WSD



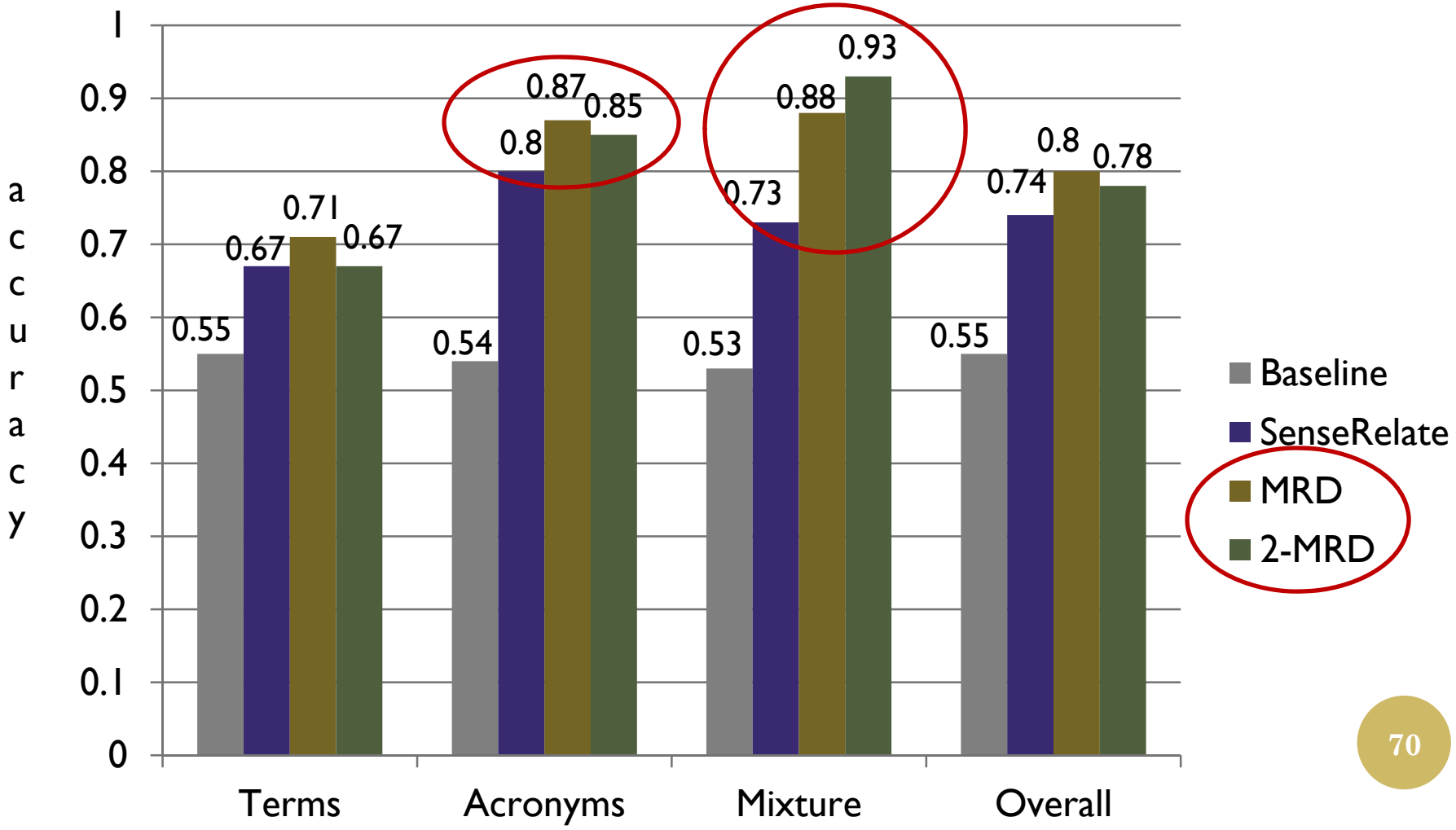
COMPARISON ACROSS SUBSETS OF MSH-WSD



COMPARISON ACROSS SUBSETS OF MSH-WSD



COMPARISON ACROSS SUBSETS OF MSH-WSD



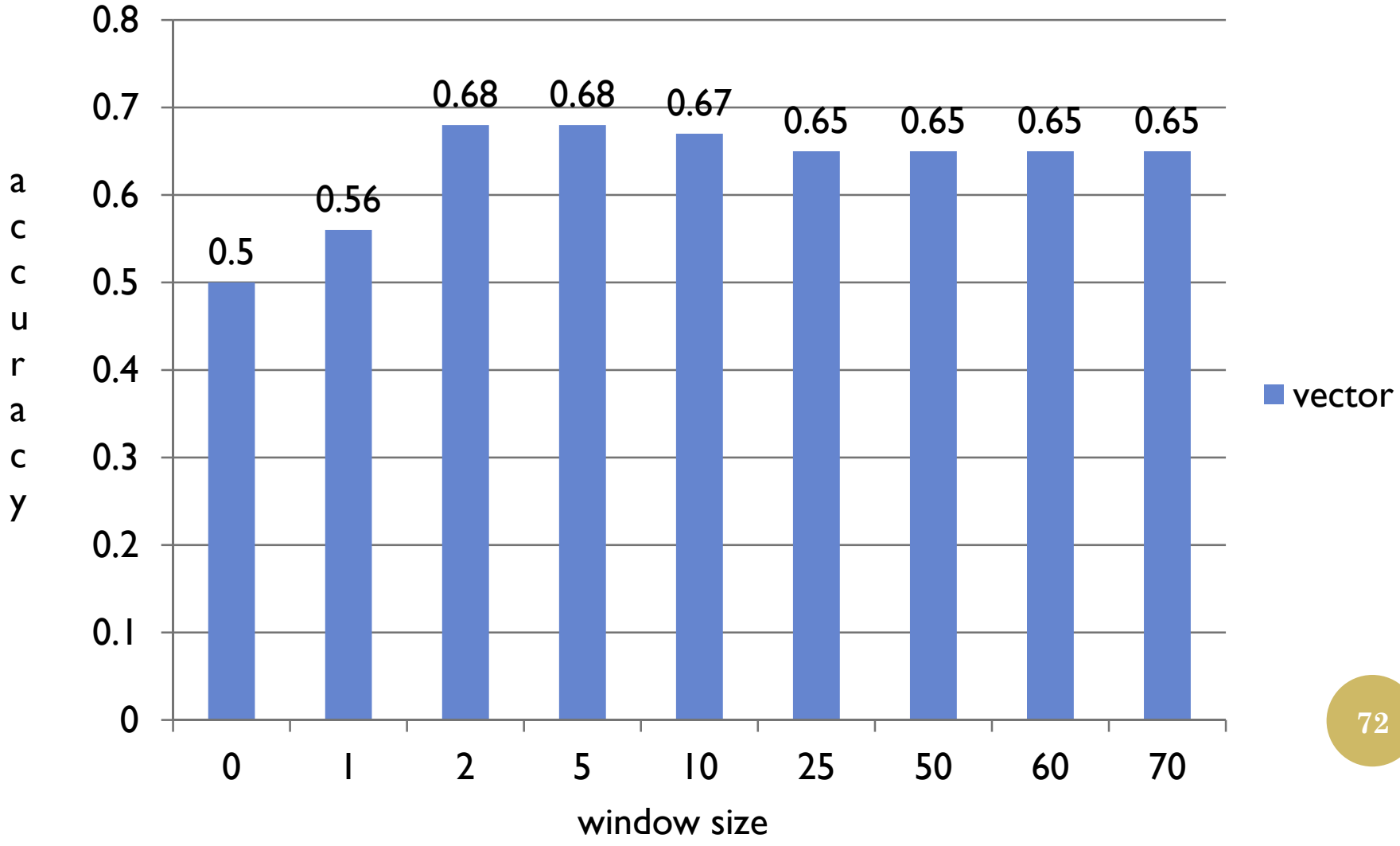
WINDOW SIZES

- Use the terms surrounding the target word within a specified window: 1, 2, 5, 10, 25, 50, 60, 70

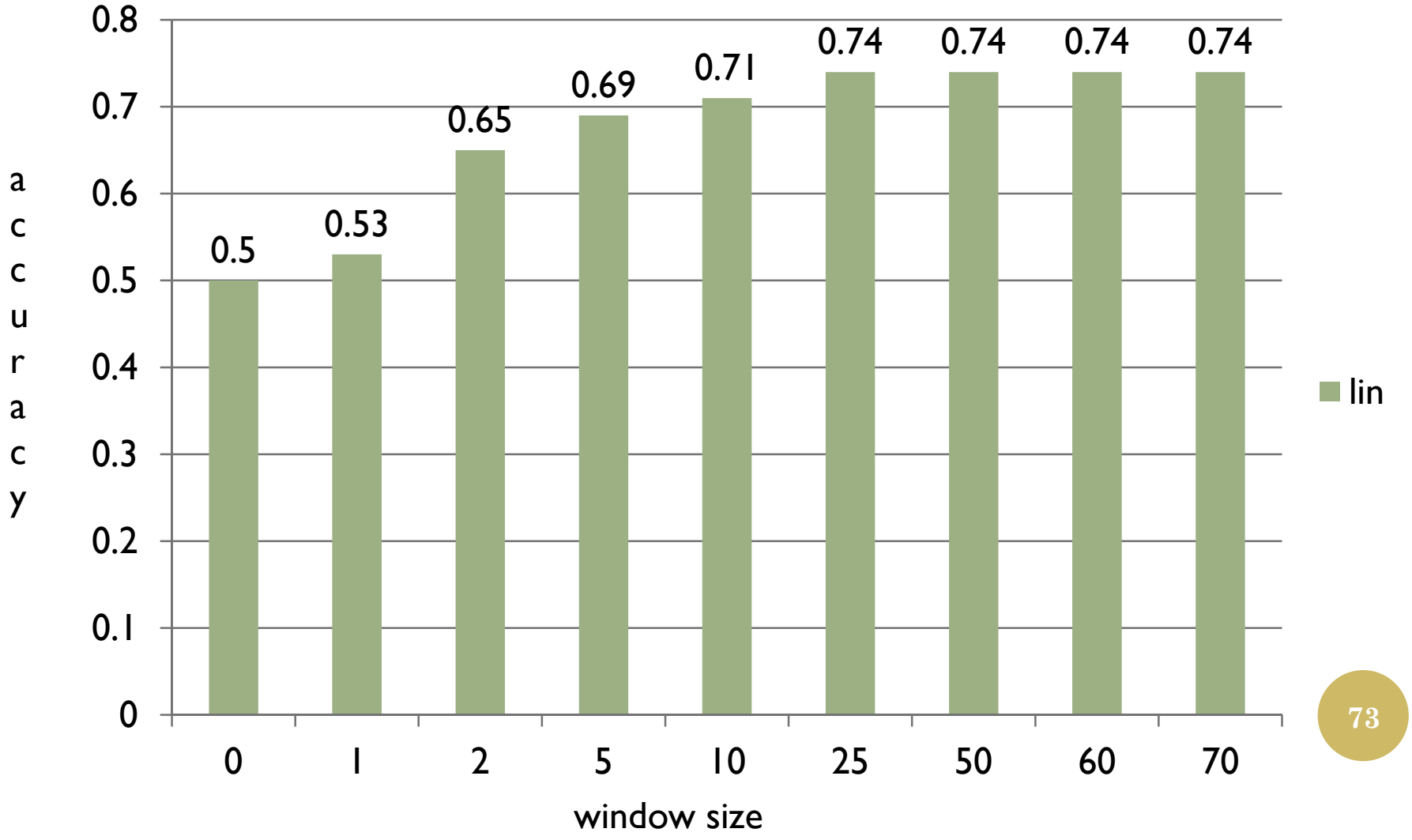
WINDOW SIZE = 2

Busprione attenuates tolerance to morphine in mice with skin_cancer

COMPARISON OF WINDOW SIZES FOR VECTOR



COMPARISON OF WINDOW SIZES FOR LIN



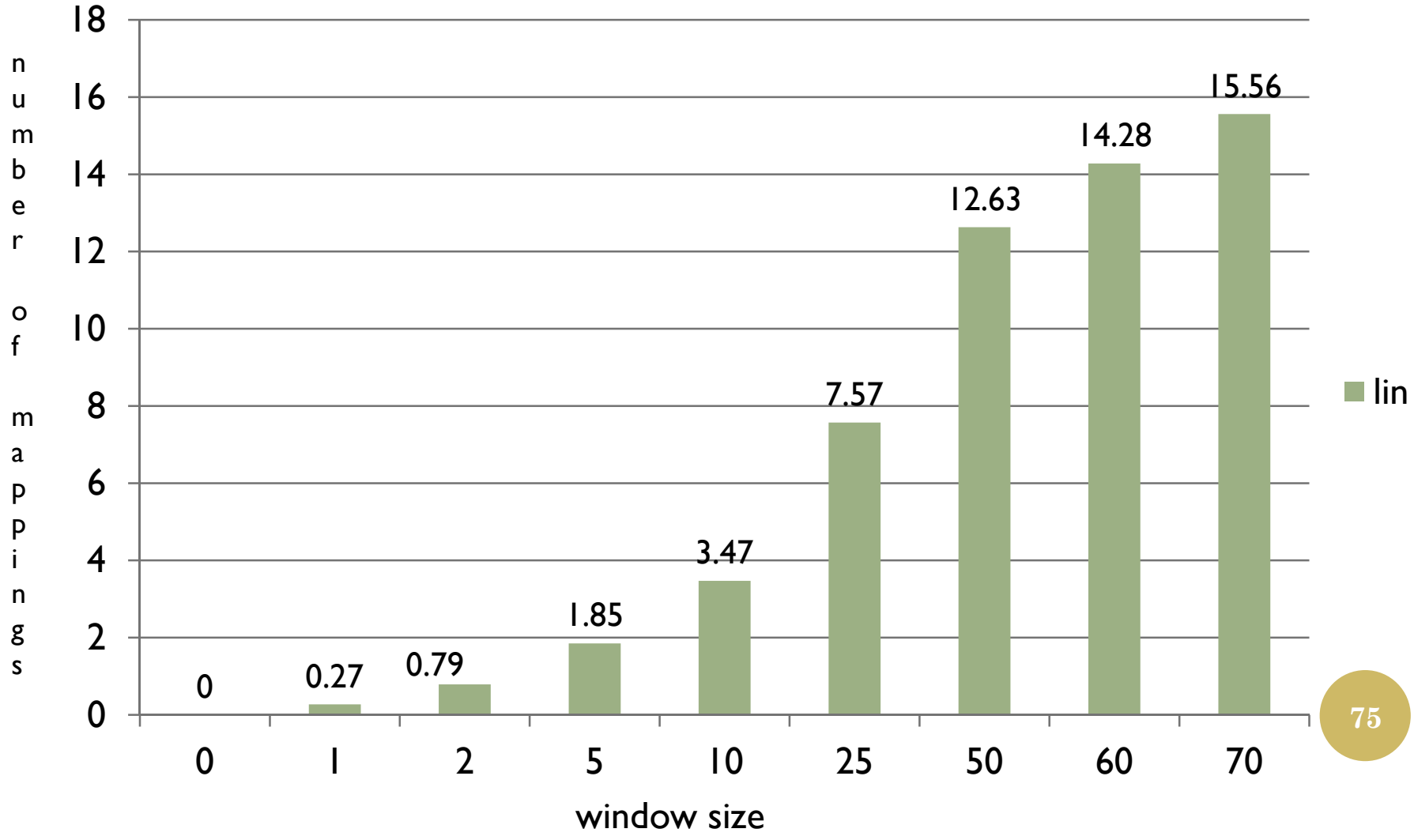
SURROUNDING TERMS

Not all terms have a concept in the UMLS

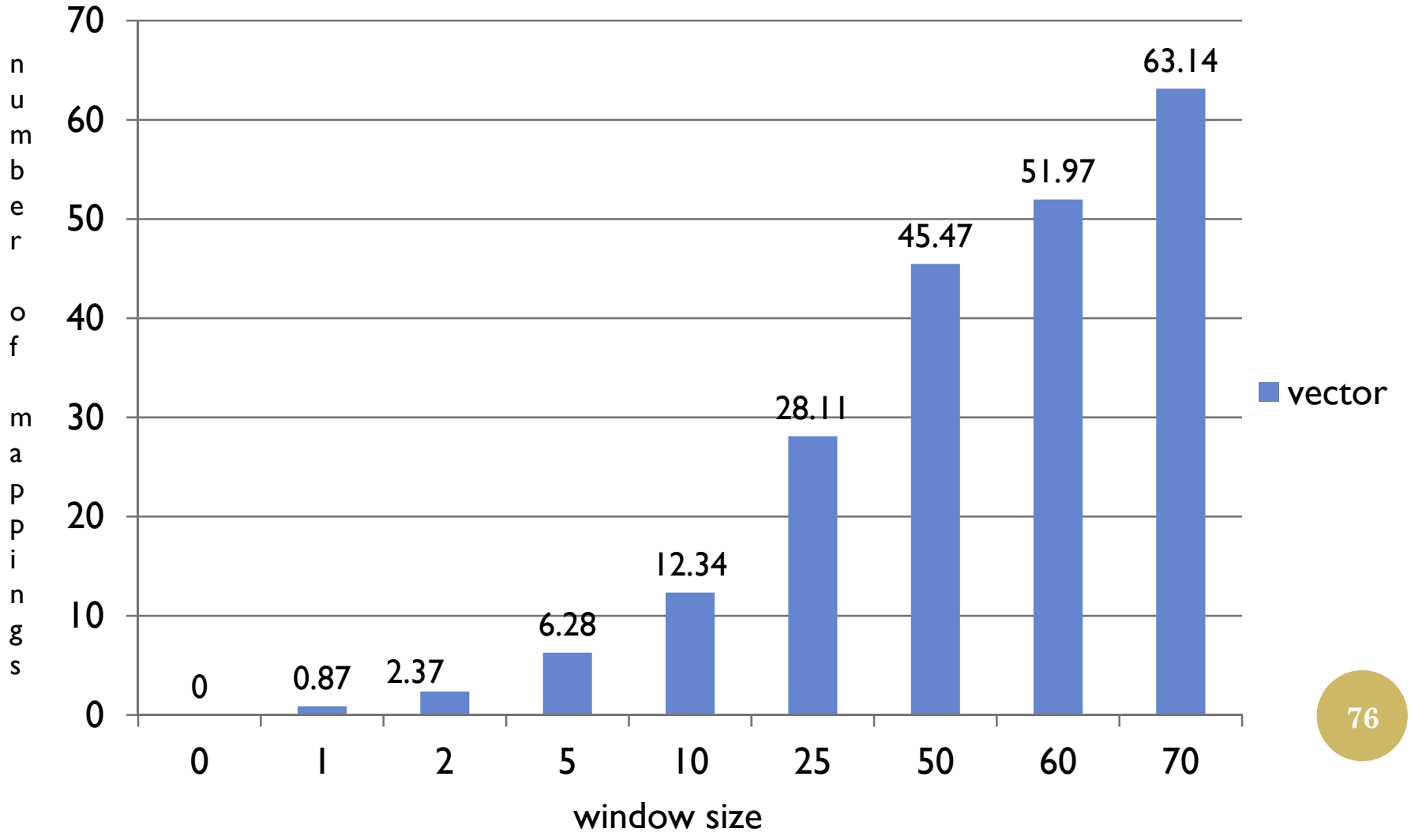
therefore

Not all surrounding terms in the window mapped to CUIs

WINDOW SIZES VERSUS MAPPED TERMS



WINDOW SIZES VERSUS MAPPED TERMS



OBJECTIVE #1

Develop and evaluate a method than can

disambiguate terms in biomedical text by

exploiting similarity and relatedness information

extrapolated from the Unified Medical Language System

- UMLS::SenseRelate statistically significantly higher disambiguation accuracy than the baseline
- On par with previous unsupervised methods

OBJECTIVE #2

Evaluate the efficacy of similarity measures and relatedness measures for WSD

- There is no statistically significant difference between the accuracies obtained by the IC-based measures
- There is a statistically significant difference between:
  - IC-based measures and the path-based measures
  - IC-based measures and relatedness measure

TAKE HOME MESSAGE:

An ambiguous word is often used in the sense that is most similar to the sense of the concepts of the terms that surround it

RESOURCES

- Software:
  - UMLS::SenseRelate
    - http://search.cpan.org/dist/UMLS-SenseRelate/
  - UMLS::Similarity
    - http://search.cpan.org/dist/UMLS-Similarity/
  - UMLS::Interface
    - http://search.cpan.org/dist/UMLS-Interface/
- Data
  - MSH-WSD
    - http://wsd.nlm.nih.gov/collaboration.shtml

RESOURCES

- Software:
  - UMLS::SenseRelate
    - http://search.cpan.org/dist/UMLS-SenseRelate/
  - UMLS::Similarity
    - http://search.cpan.org/dist/UMLS-Similarity/
  - UMLS::Interface
    - http://search.cpan.org/dist/UMLS-Interface/
- Data
  - MSH-WSD
    - http://wsd.nlm.nih.gov/collaboration.shtml

THANK YOU

RESOURCES

- Software:
  - UMLS::SenseRelate
    - http://search.cpan.org/dist/UMLS-SenseRelate/
  - UMLS::Similarity
    - http://search.cpan.org/dist/UMLS-Similarity/
  - UMLS::Interface
    - http://search.cpan.org/dist/UMLS-Interface/
- Data
  - MSH-WSD
    - http://wsd.nlm.nih.gov/collaboration.shtml

QUESTIONS?

Vector Results With Data sources