Knowledge-based Method for Determining the Meaning of Ambiguous Biomedical Terms Using Information Content Measures of Similarity

Bridget McInnes
Ted Pedersen
Ying Liu
Genevieve B. Melton
Serguei Pakhomov

Objective of this work

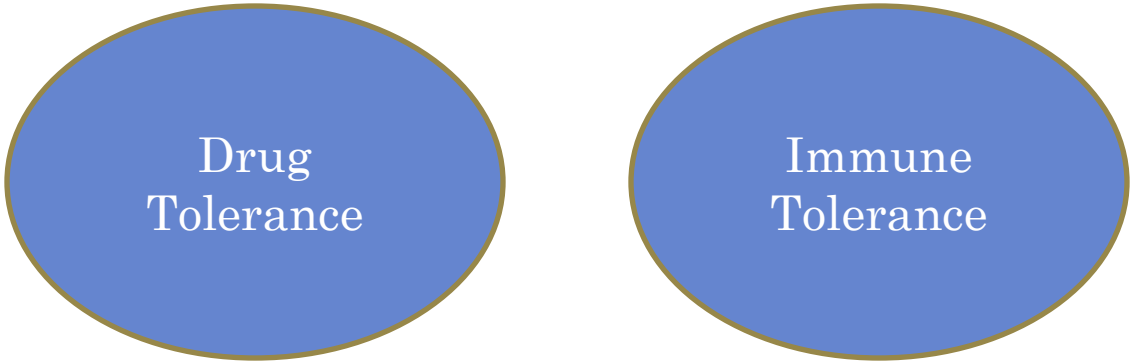
- Develop and evaluate a method than can
  disambiguate terms in biomedical text by
  exploiting similarity information
  extrapolated from the Unified Medical Language System

- Evaluate the efficacy of Information Content-based similarity measures over path-based similarity measures for Word Sense Disambiguation, WSD

Word Sense Disambiguation

Word sense disambiguation is the task of determining the appropriate sense of a term given context in which it is used.

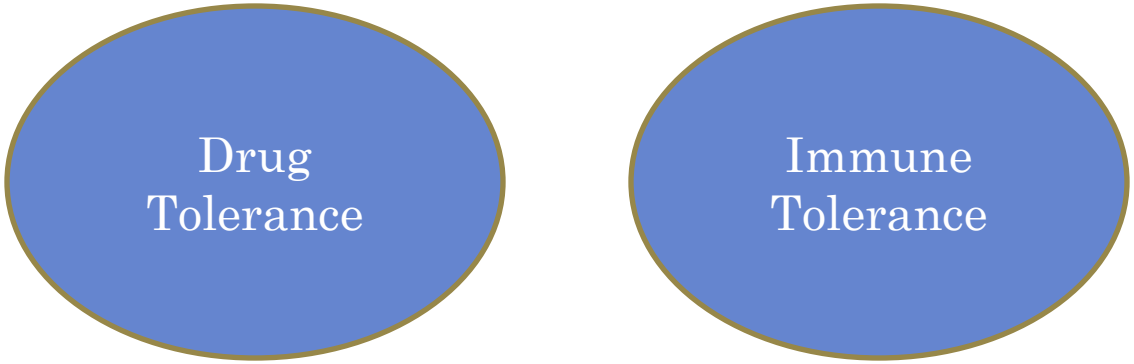
TERM: tolerance



Word Sense Disambiguation

Word sense disambiguation is the task of determining the appropriate sense of a term given context in which it is used.

Busprione attenuates tolerance to morphine in mice with skin cancer

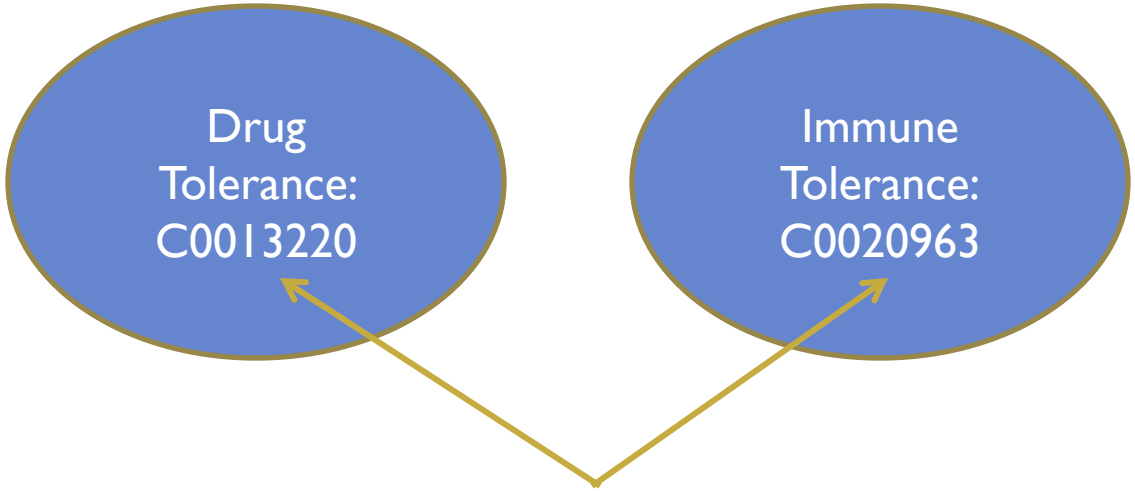


Sense inventory: Unified Medical Language System

- Unified Medical Language Sources (UMLS)
  - Semantic Network
  - Metathesaurus
    - ~1.7 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



Concept Unique Identifiers: CUIs

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

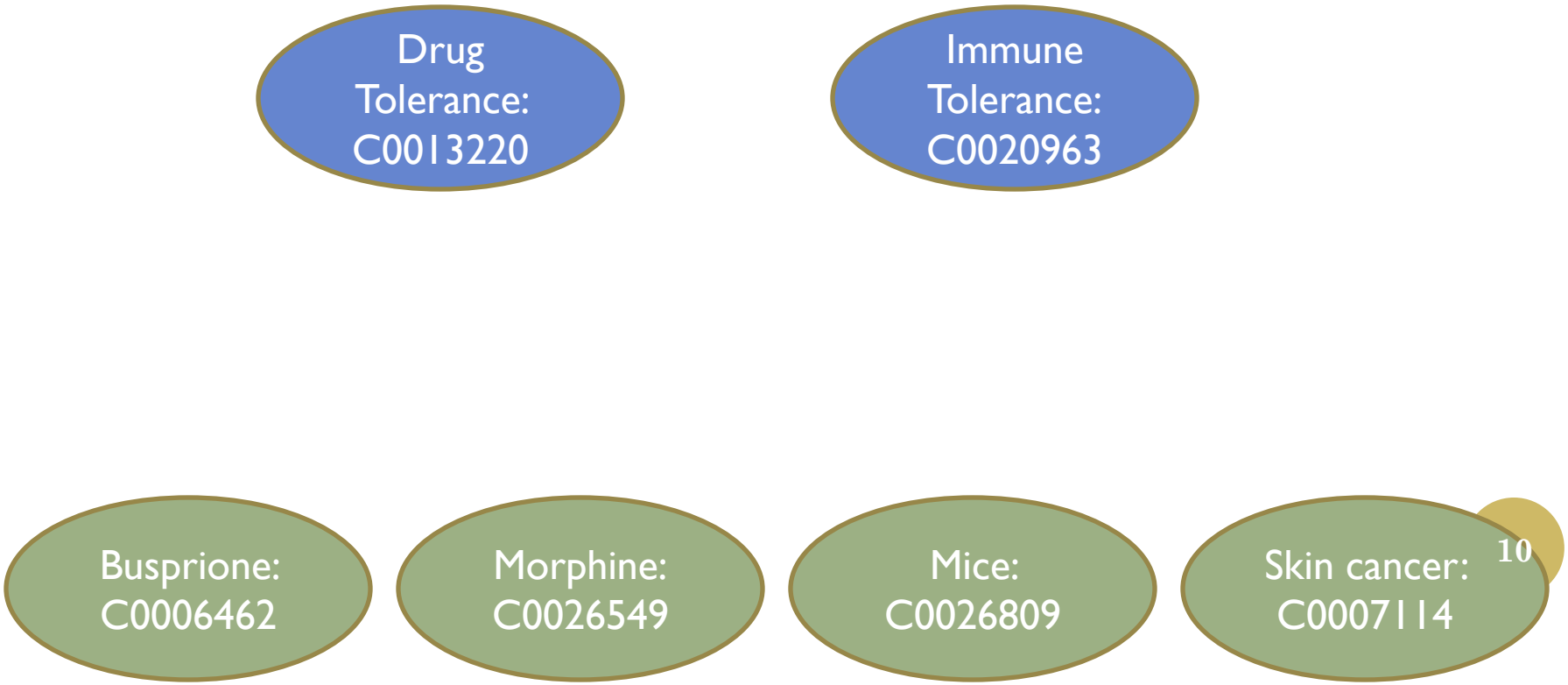
Busprione attenuates tolerance to morphine in mice with skin cancer

SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer

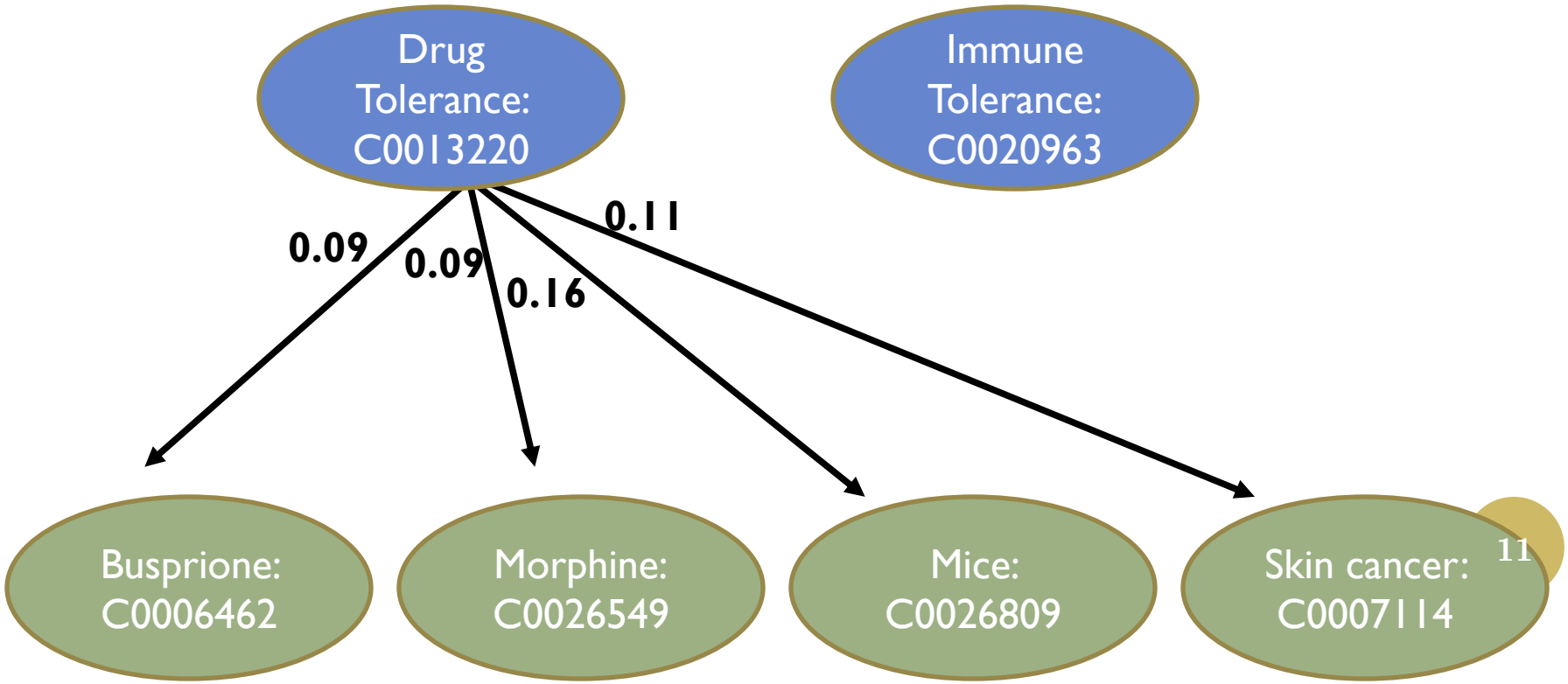
SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer



SENSERELATE EXAMPLE

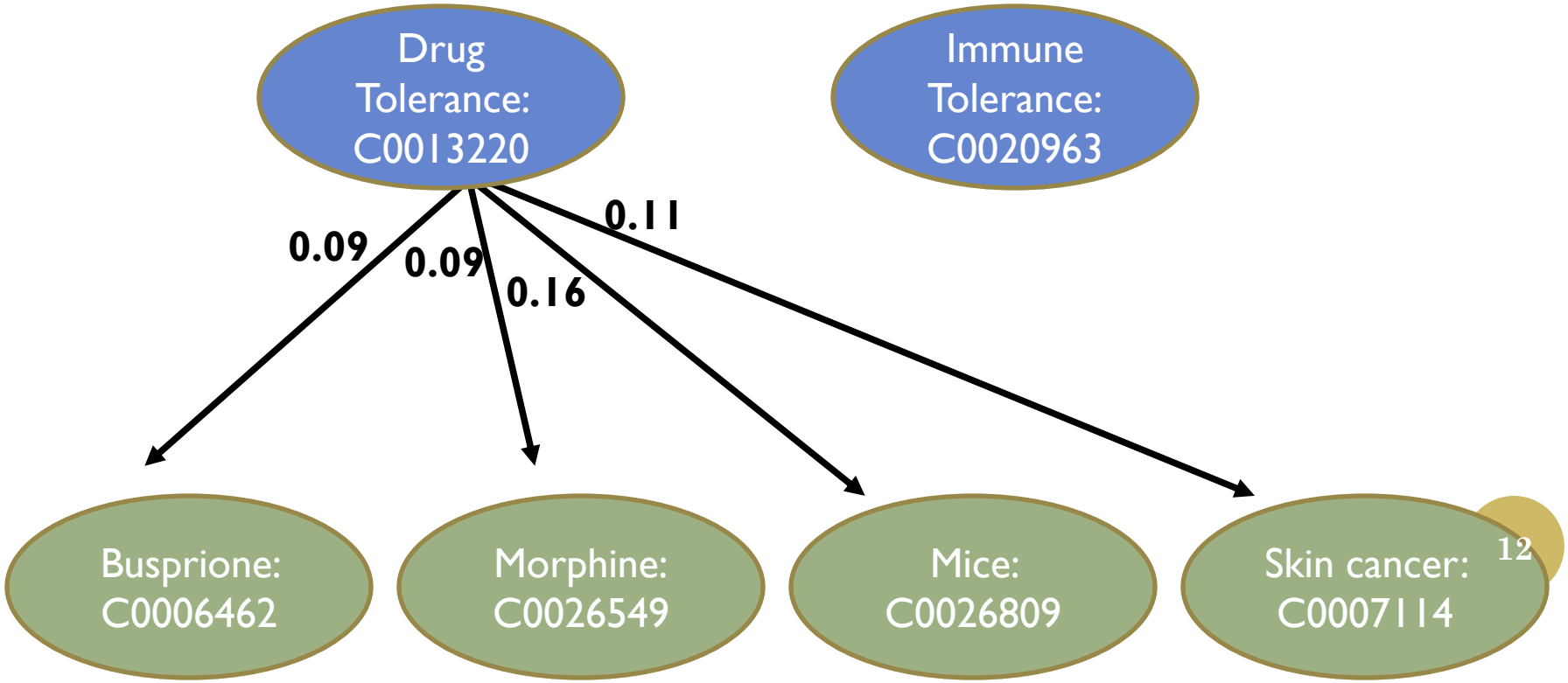
Busprione attenuates tolerance to morphine in mice with skin cancer



SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer

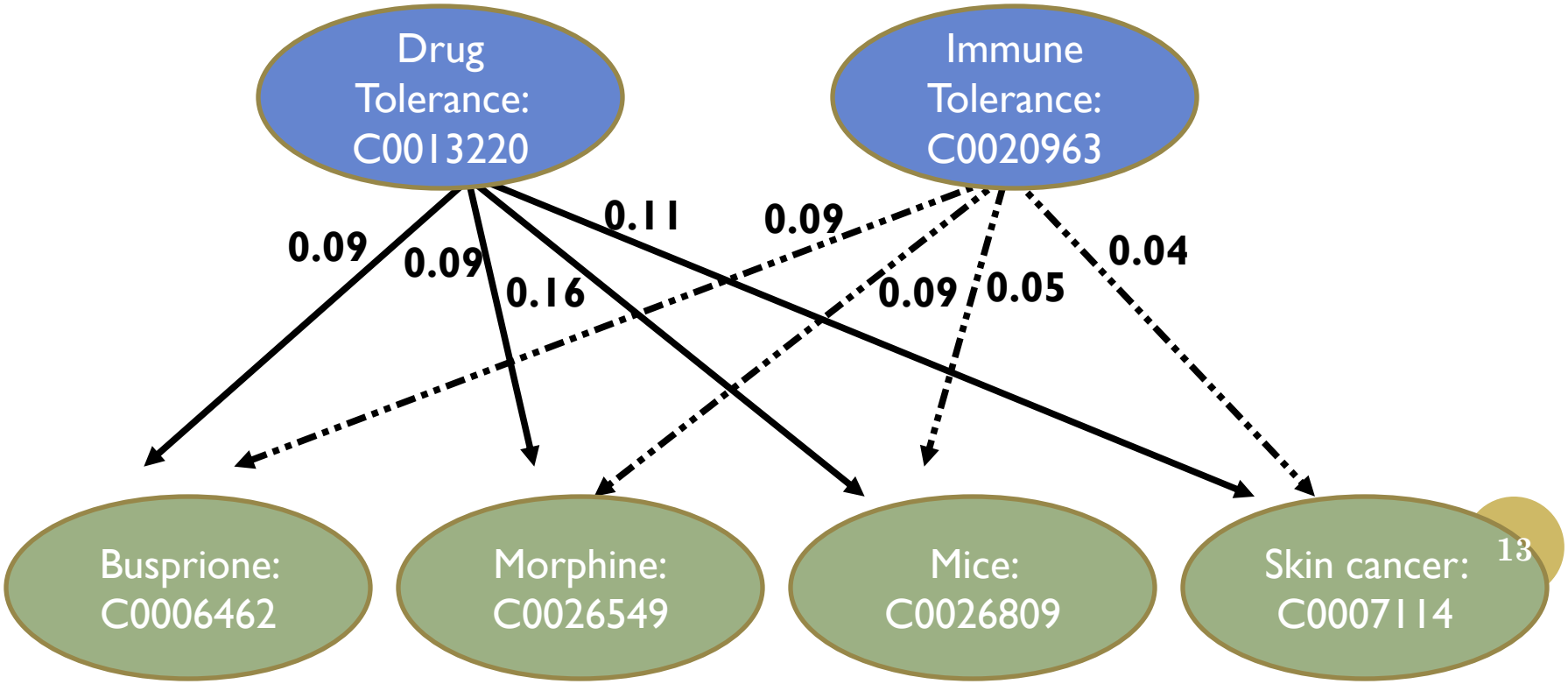
Drug Tolerance
Score = 0.09 + 0.09 + 0.16 + 0.11 = 0.45



SENSERELATE EXAMPLE

Busprione attenuates tolerance to morphine in mice with skin cancer

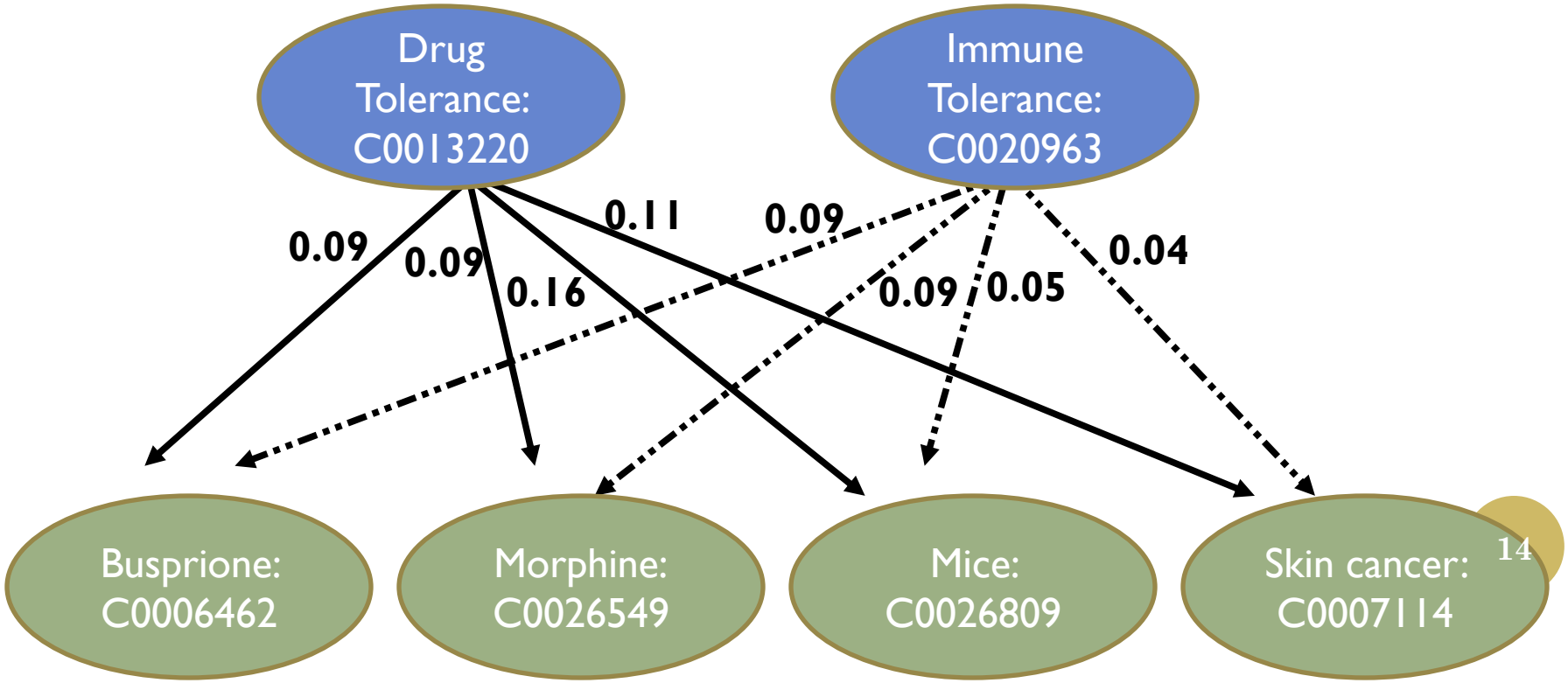
Drug Tolerance
Score = 0.09 + 0.09 + 0.16 + 0.11 = 0.45



SENSERELATE EXAMPLE

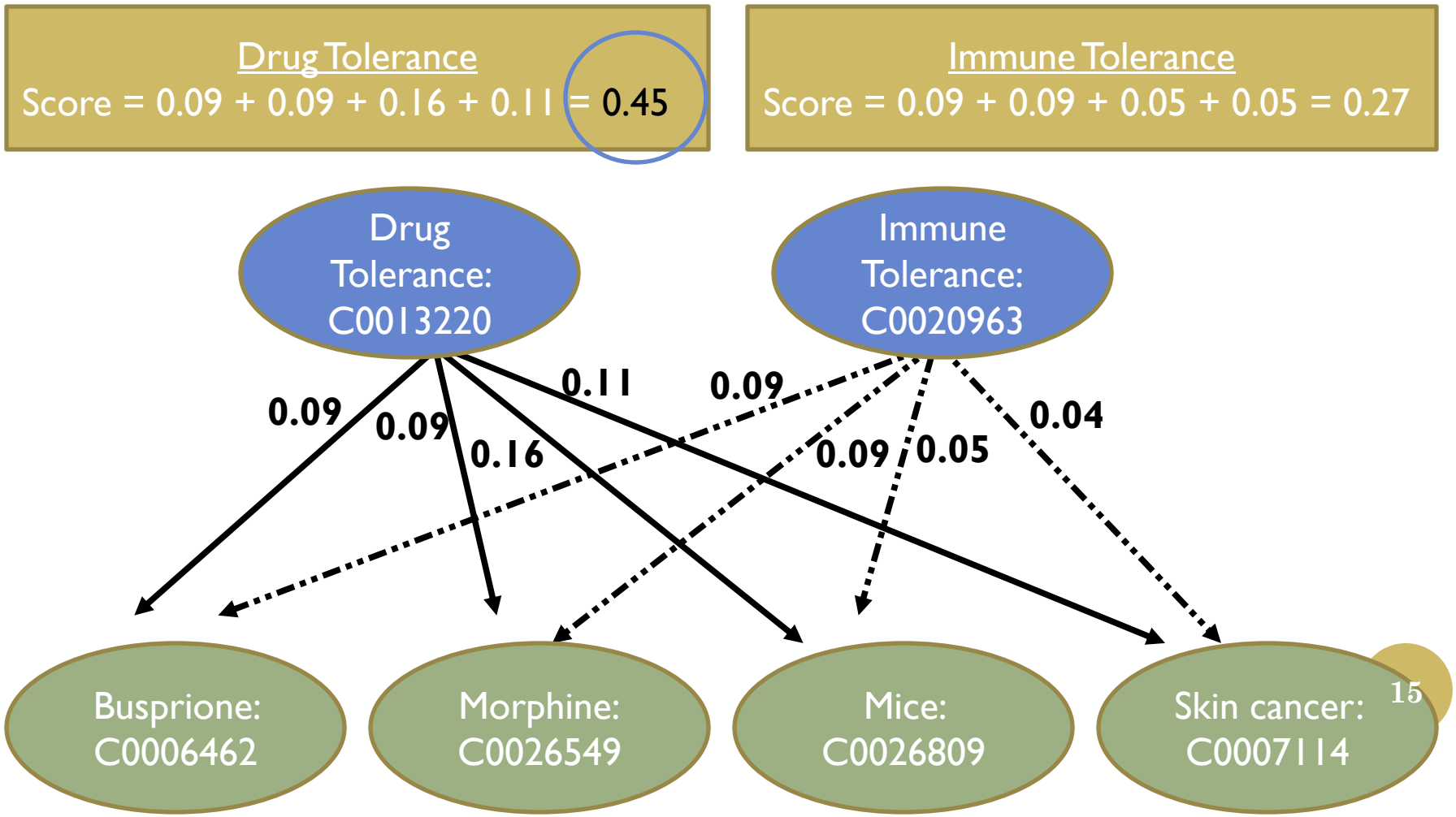
Busprione attenuates tolerance to morphine in mice with skin cancer

Drug Tolerance	Immune Tolerance
Score = 0.09 + 0.09 + 0.16 + 0.11 = 0.45	Score = 0.09 + 0.09 + 0.05 + 0.05 = 0.27



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

SenseRelate Components

- Identifying the concepts of surrounding terms
- Calculating semantic similarity

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

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

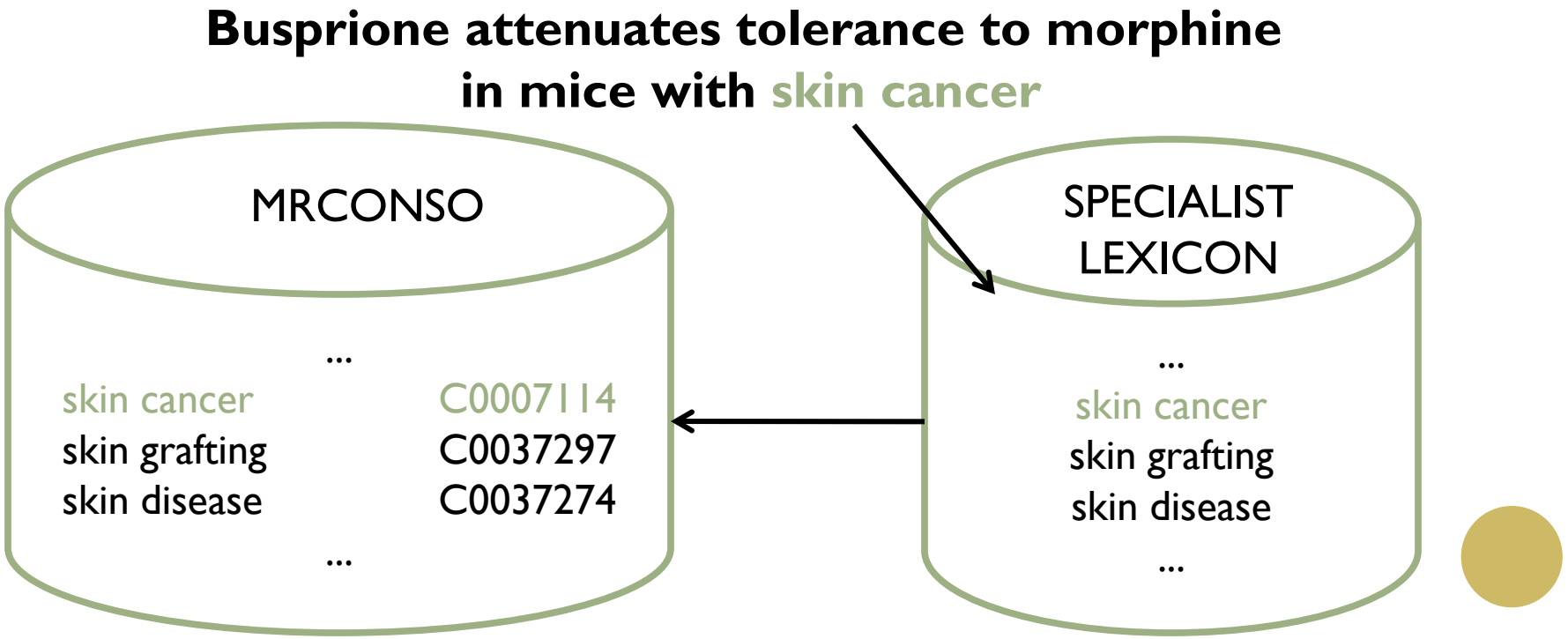
Identifying the concepts of the surrounding terms

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

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



Semantic Similarity Measures

- Path-based measures
  - Path
  - Wu and Palmer
  - Leacock and Chodorow
  - Ngyuen and Al-Mubaid
- Information content (IC)-based measures
  - Resnik
  - Lin
  - Jiang and Conrath

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

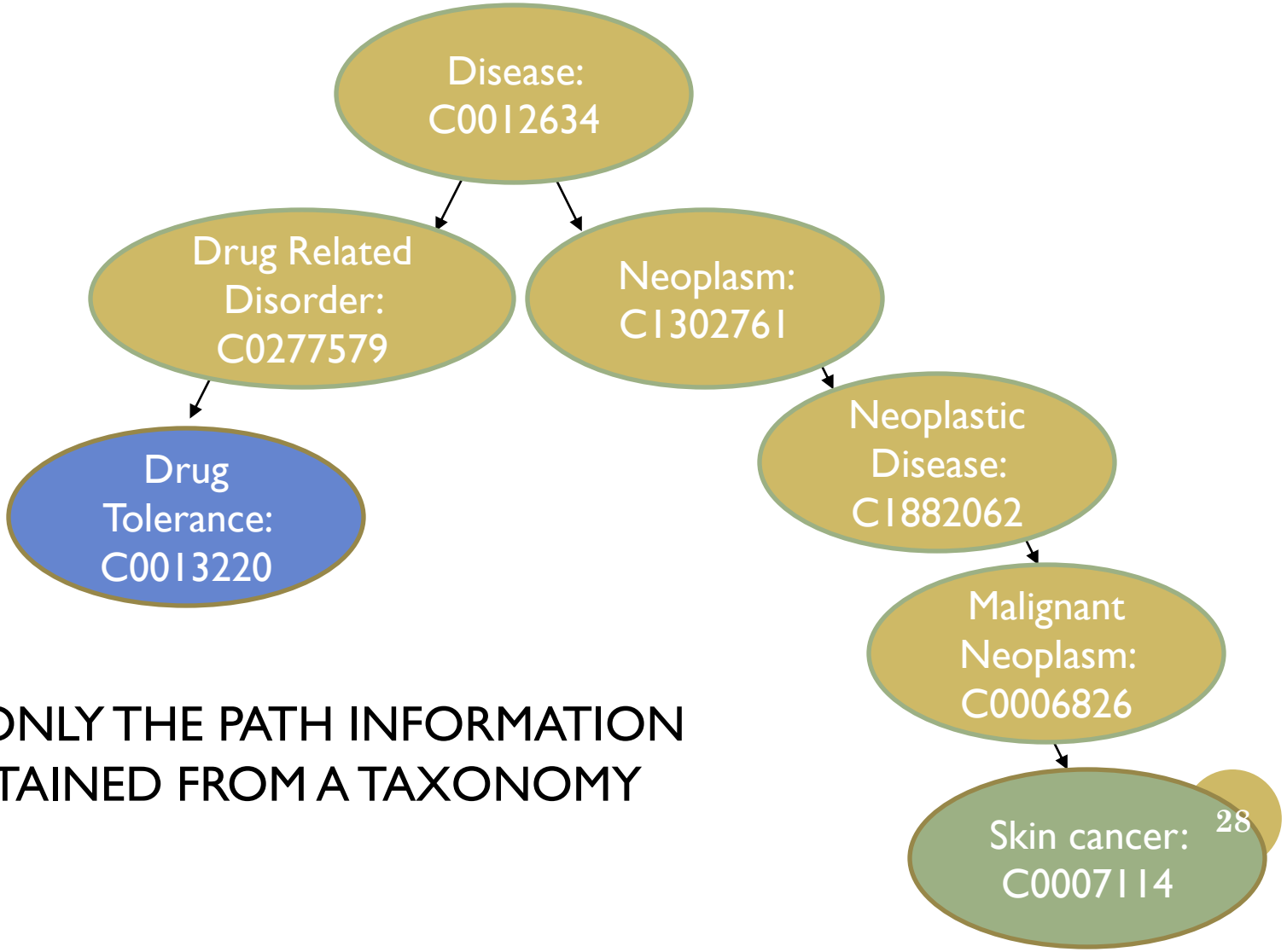
PATH-BASED SIMILARITY MEASURES

- 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

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

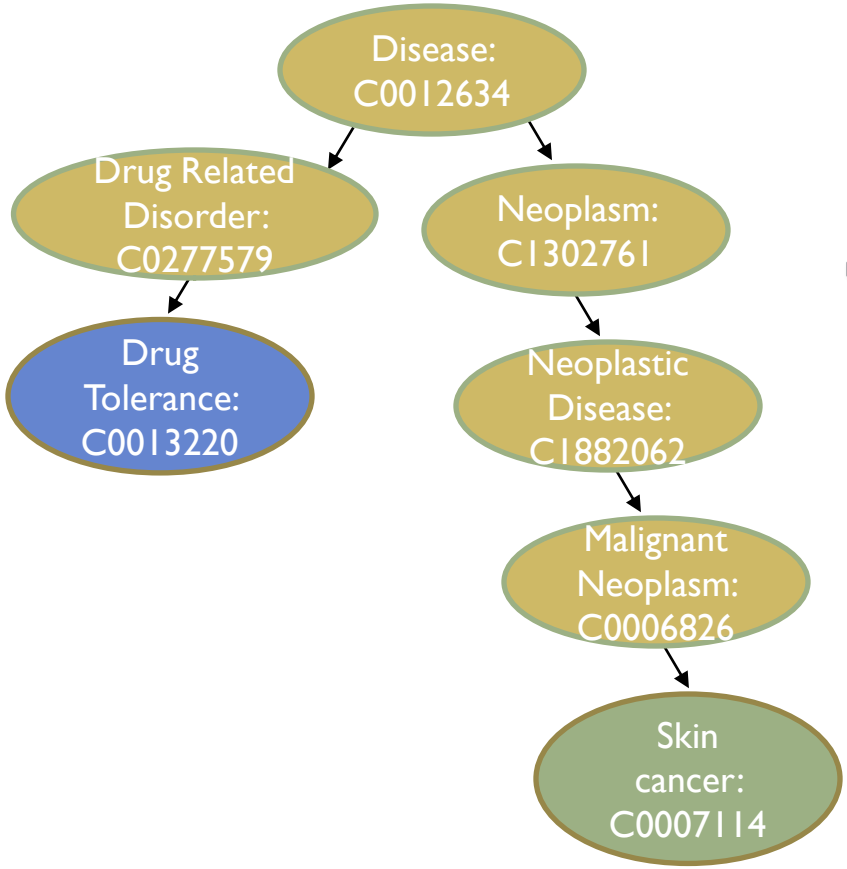
- Incorporate the probability of the concepts
  - IC = -log(P(concept)
- Resnik, 1995
  - sim(c1,c2) = IC(LCS(c2,c2))
- Jiang and Conrath, 1997
  - sim(c1,c2) = 1 / (IC(c1)+IC(c2) – 2* IC(LCS(c1,c2))

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))

IC-BASED SIMILARITY MEASURES

PATH INFORMATION



PROBABILITY OF CONCEPTS

EXTERNAL CORPUS

Experimental Framework

- Use open-source UMLS::Similarity package to obtain the similarity between the terms and possible senses in the SenseRelate algorithm

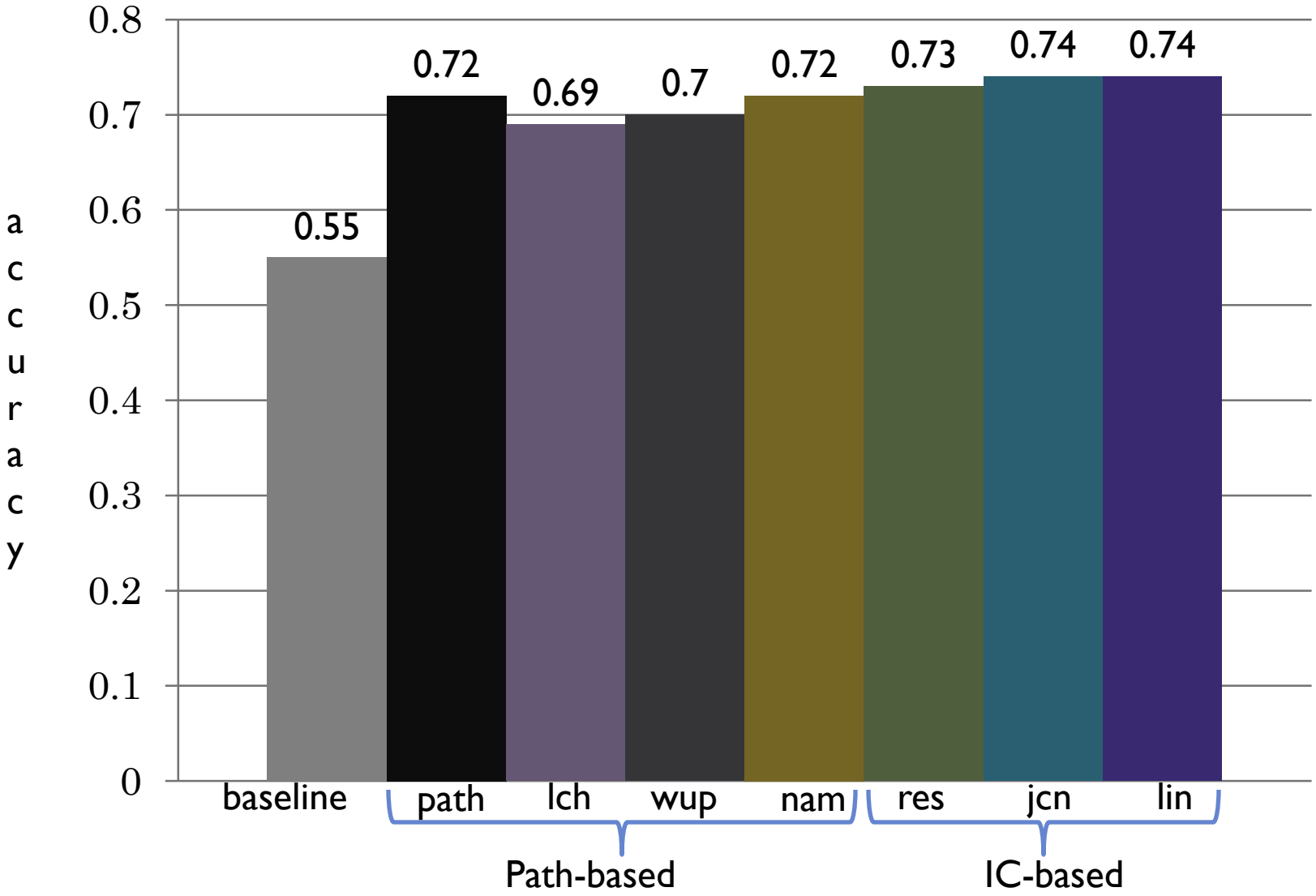
- Path information: parent/child relations in MSH source

- 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

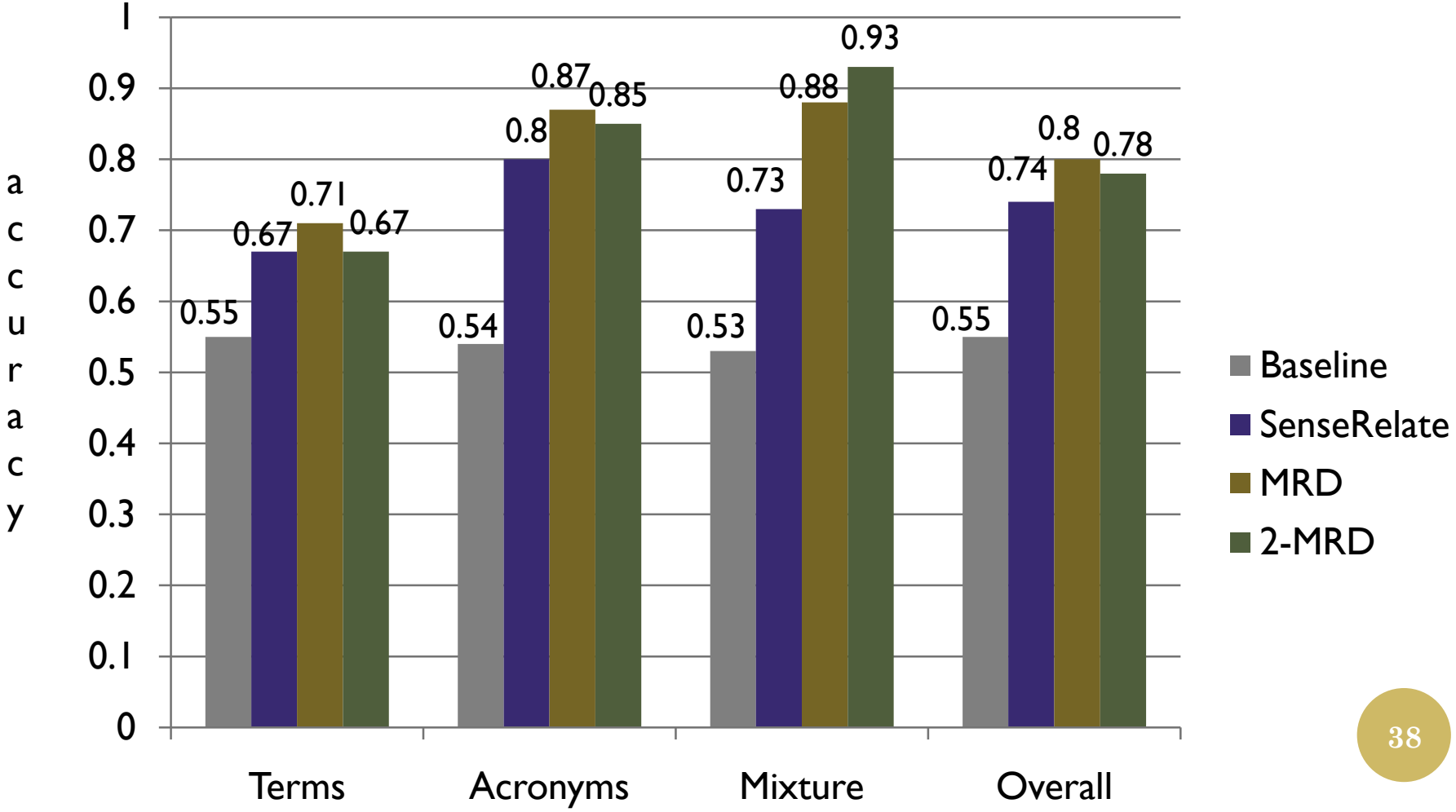
Evaluation Data: MSH WSD

- 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%

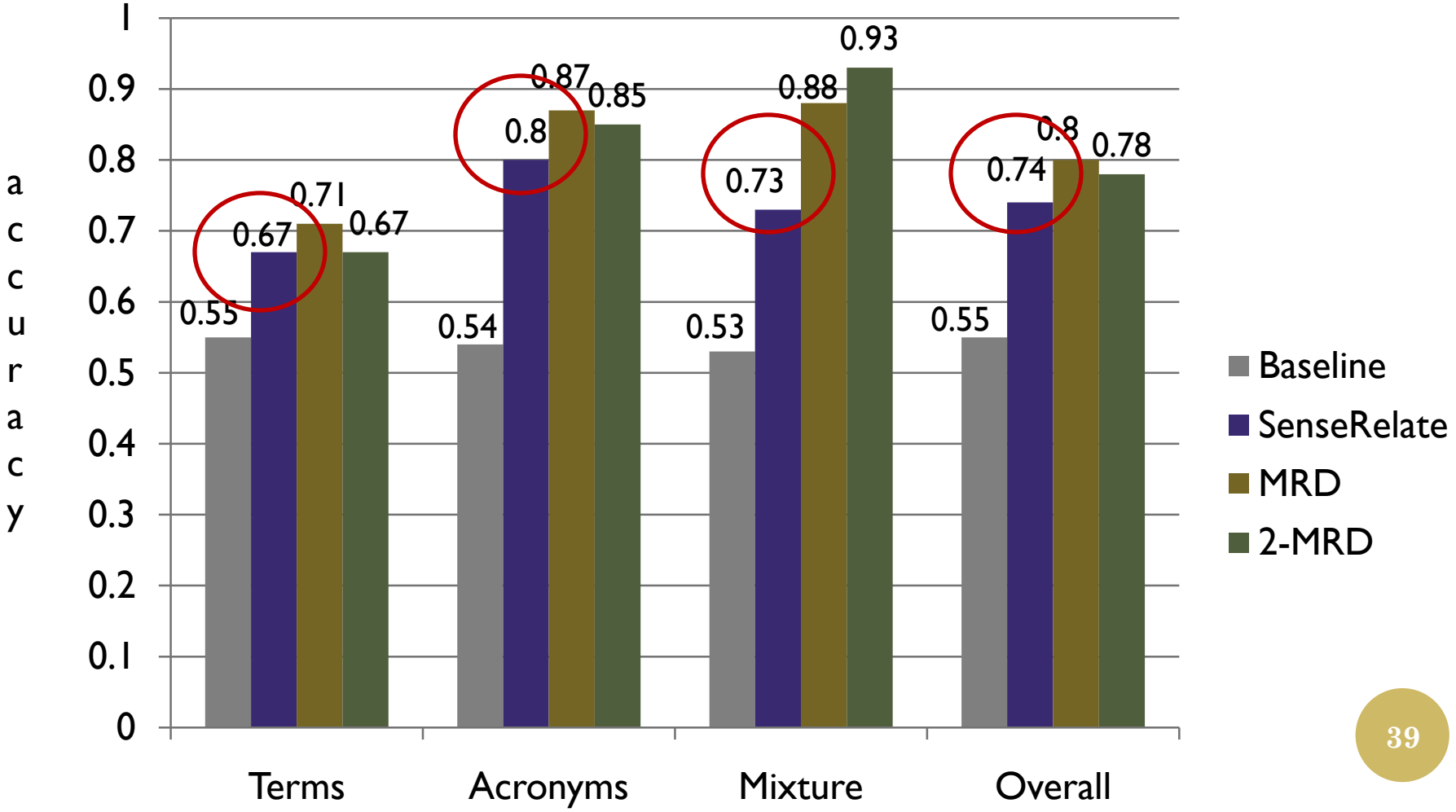
RESULTS


accuracy
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.55
0.72
0.69
0.7
0.72
0.73
0.74
0.74
baseline
path
lch
wup
nam
res
jcn
lin
Path-based
IC-based

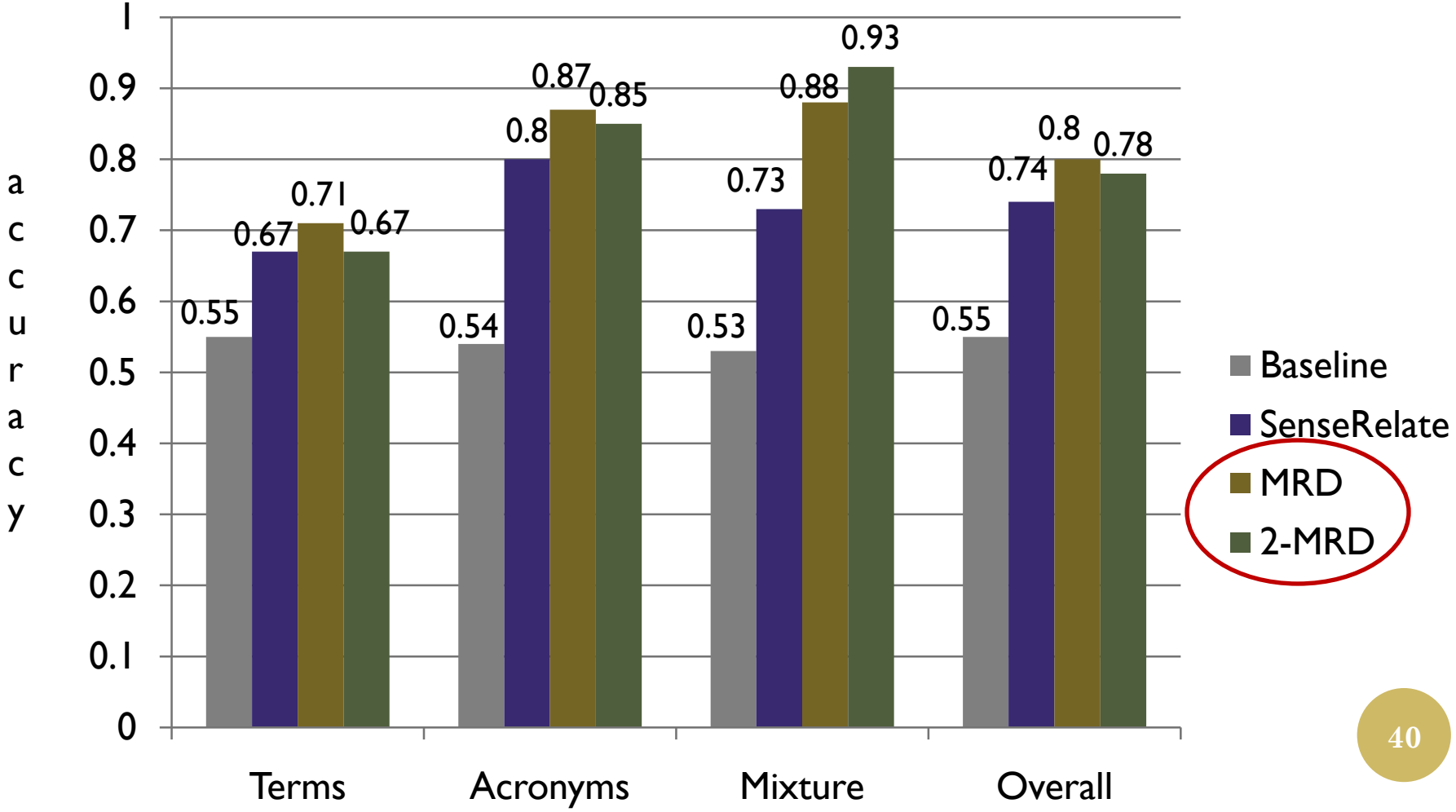
Comparison across subsets of MSH-WSD



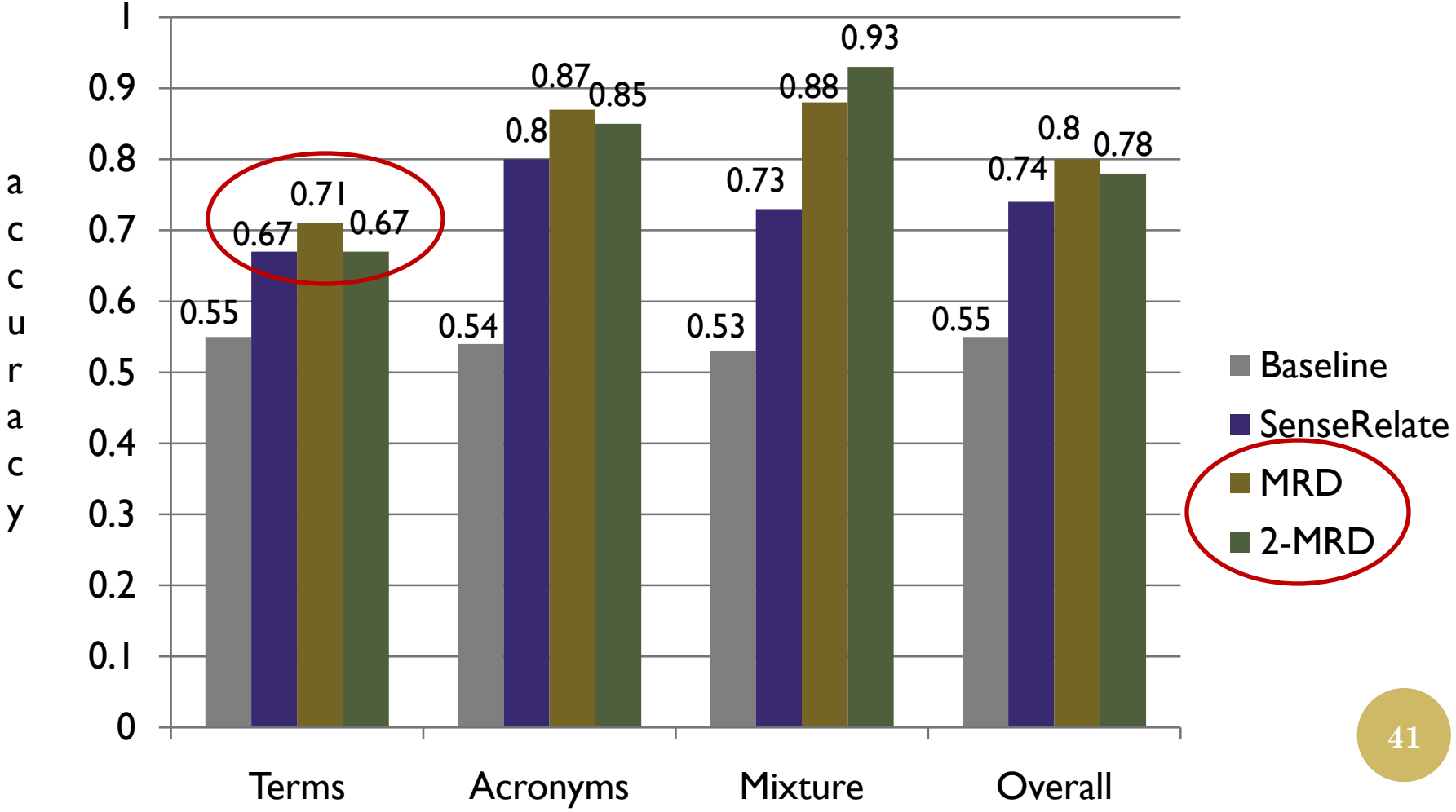
Comparison across subsets of MSH-WSD



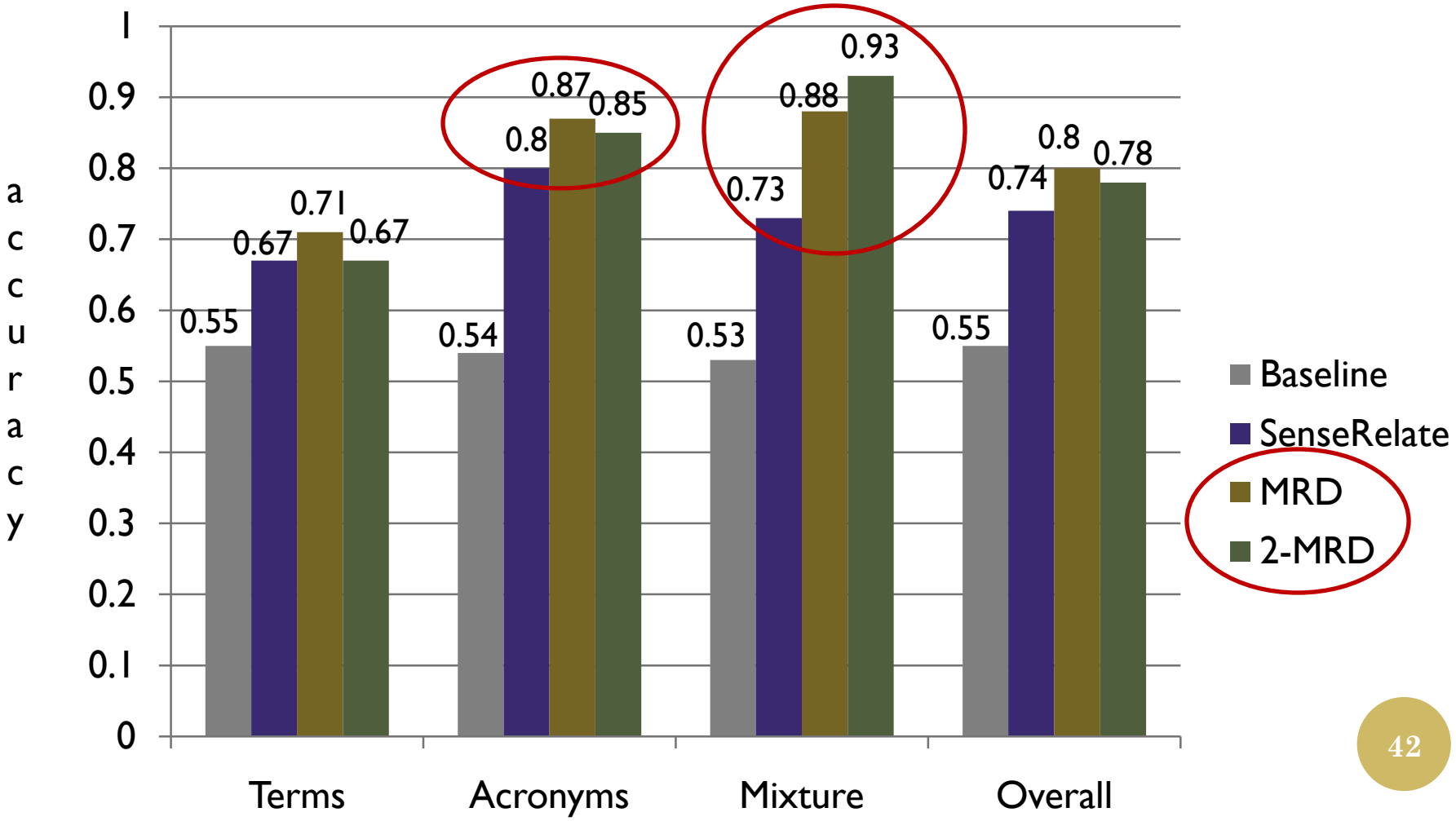
Comparison across subsets of MSH-WSD



Comparison across subsets of MSH-WSD



Comparison across subsets of MSH-WSD



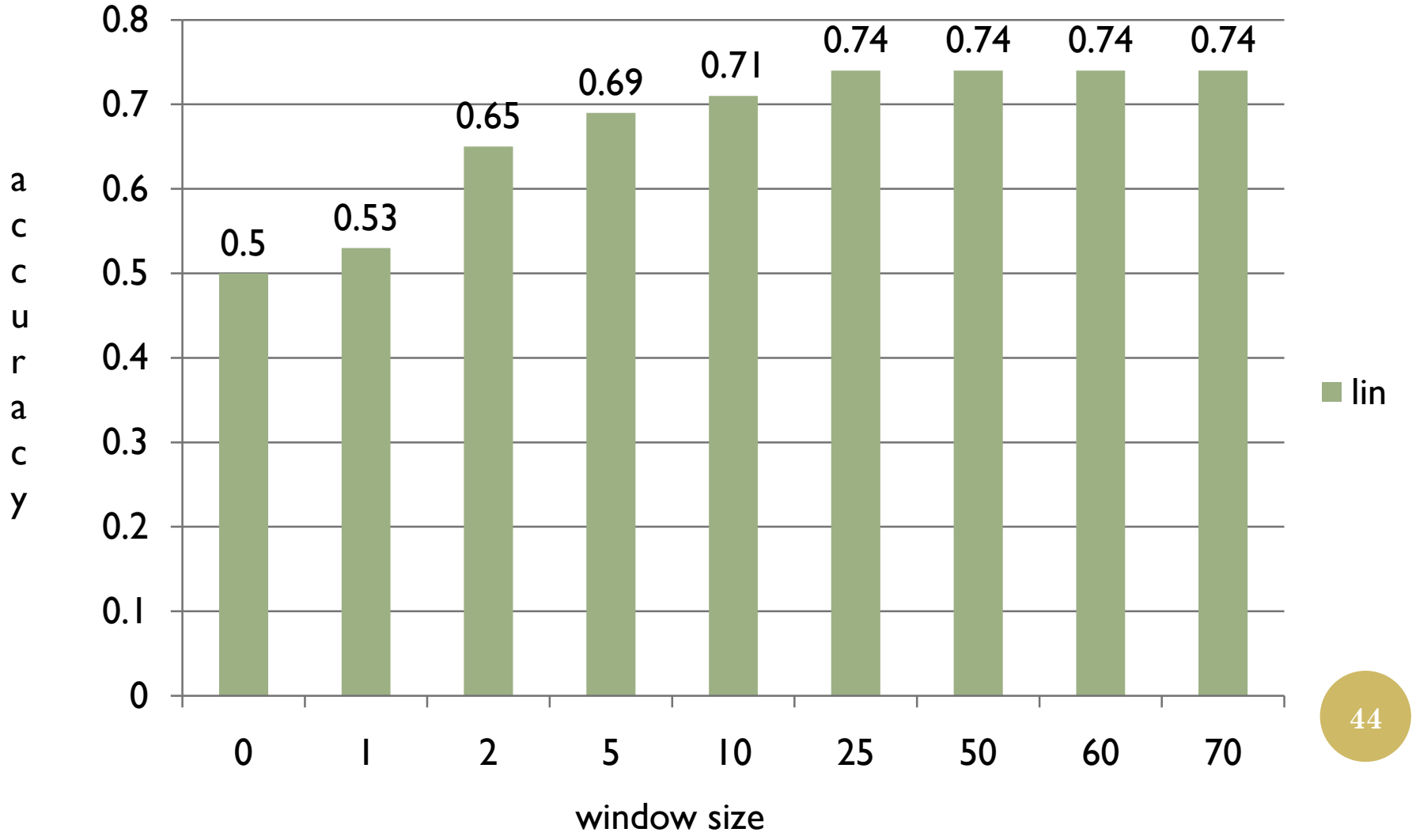
WINDOW SIZES

- 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 lin



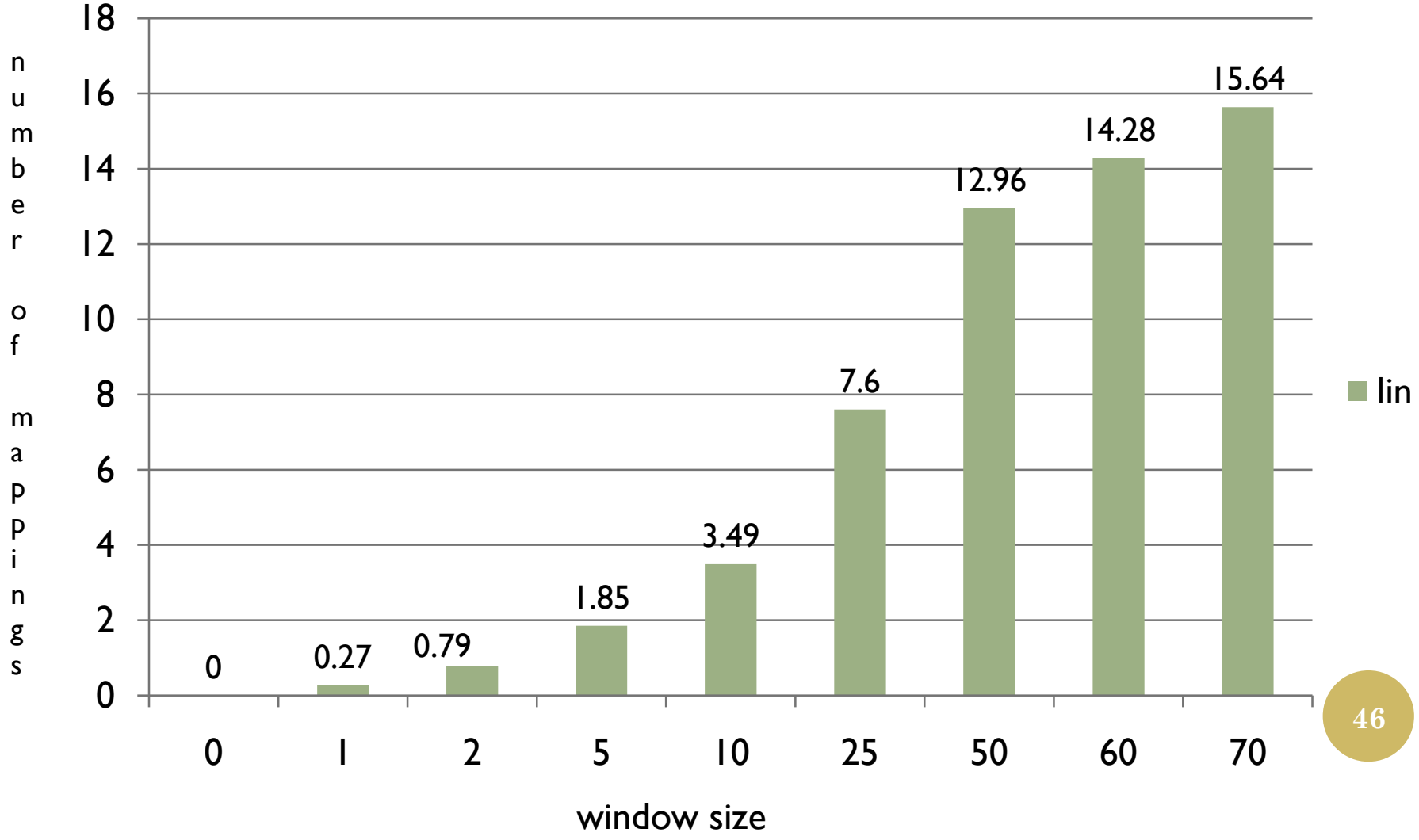
SURROUNDING TERMS

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



Future Work: Mapping Terms

- Currently looking at mapping the terms to CUIs using information from the concept mapping system MetaMap
  - Obtain the terms from MetaMap and do a dictionary look up in MRCONSO
    - Hypothesis – the terms obtained by MetaMap are more accurate than using the SPECIALIST Lexicon
  - Obtain the CUIs from MetaMap
    - Hypothesis – the CUIs obtained by MetaMap will be more accurate than the dictionary look-up

Objective #1

Develop and evaluate a method than can disambiguate terms in biomedical text by exploiting similarity information extrapolated from the UMLS

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

Objective #2

Evaluate the efficacy of IC-based similarity measures over path-based measures on a secondary task

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

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/
- 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/
- 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/
- Data
  - MSH-WSD
    - http://wsd.nlm.nih.gov/collaboration.shtml

QUESTIONS?