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

Drug Tolerance

Immune 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
Busprione attenuates **tolerance** to morphine in mice with skin cancer

**Concept Unique Identifiers: CUIs**

**Drug Tolerance:** C0013220

**Immune Tolerance:** C0020963
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
SENSE RELATE EXAMPLE

Busprione attenuates **tolerance** to morphine in mice with skin cancer
**SENSE RELATE EXAMPLE**

**Busprione attenuates tolerance to morphine in mice with skin cancer**

- **Drug Tolerance:** C0013220
- **Immune Tolerance:** C0020963
**SenseRelate Example**

*Busprione attenuates tolerance to morphine in mice with skin cancer*

- **Drug Tolerance:** Busprione: C0006462
- **Immune Tolerance:** C0020963
- **Mice:** C0026809
- **Skin cancer:** C0007114
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

- **Drug Tolerance:** C0013220
  - 0.09
  - 0.09
  - 0.16
  - 0.11

- **Immune Tolerance:** C0020963

- **Busprione:** C0006462
- **Morphine:** C0026549
- **Mice:** C0026809
- **Skin cancer:** C0007114
Busprione attenuates **tolerance** to morphine in mice with skin cancer

### Drug Tolerance Score

\[
0.09 + 0.09 + 0.16 + 0.11 = 0.45
\]
Busprione attenuates tolerance to morphine in mice with skin cancer

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

**Immune Tolerance**
Score = 0.09 + 0.09 + 0.05 + 0.05 = 0.27
Busprione attenuates tolerance to morphine in mice with skin cancer

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

Immune Tolerance Score = 0.09 + 0.09 + 0.05 + 0.05 = 0.27
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
  - \( \text{sim}(c_1, c_2) = \frac{1}{\text{minpath}(c_2, c_2)} \)
    - where \( \text{minpath} \) is the shortest path between the two concepts
PATH-BASED SIMILARITY MEASURES

- Use only the path information obtained from a taxonomy

- Path measure
  - \( \text{sim}(c_1,c_2) = \frac{1}{\text{minpath}(c_2,c_2)} \)
    - where minpath is the shortest path between the two concepts

- Wu and Palmer, 1994
  - \( \text{sim}(c_1,c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_2,c_2))}{\text{depth}(c_1)+\text{depth}(c_2)} \)
    - 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
  - \( \text{sim}(c_1, c_2) = \frac{1}{\text{minpath}(c_2, c_2)} \)
    - where minpath is the shortest path between the two concepts

- Wu and Palmer, 1994
  - \( \text{sim}(c_1, c_2) = \frac{(2 \times \text{depth(LCS(c_2, c_2)))}}{(\text{depth}(c_1) + \text{depth}(c_2))} \)
    - where LCS is the least common subsumer of the two concepts

- Leacock and Chodorow, 1998
  - \( \text{sim}(c_1, c_2) = -\log\left( \frac{\text{minpath}(c_1, c_2)}{2D} \right) \)
    - where D is the total depth of the taxonomy
PATH-BASED SIMILARITY MEASURES

- Use only the path information obtained from a taxonomy

- Path measure
  - \( \text{sim}(c_1, c_2) = \frac{1}{\text{minpath}(c_2, c_2)} \)
    - where \( \text{minpath} \) is the shortest path between the two concepts

- Leacock and Chodorow, 1998
  - \( \text{sim}(c_1, c_2) = -\log \left( \frac{\text{minpath}(c_1, c_2)}{2D} \right) \)
    - where \( D \) is the total depth of the taxonomy

- Wu and Palmer, 1994
  - \( \text{sim}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_2, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \)
    - where \( \text{LCS} \) is the least common subsumer of the two concepts

- Nyguen and Al-Mubaid, 2006
  - \( \text{sim}(c_1, c_2) = \log \left( 2 + \text{minpath}(c_1, c_2) - 1 \right) \times \left( D - \text{depth}(\text{LCS}(c_1, c_2)) \right) \)
PATH-BASED SIMILARITY MEASURES

Disease: C0012634

Drug Related Disorder: C0277579

Drug Tolerance: C0013220

Neoplasm: C1302761

Neoplastic Disease: C1882062

Malignant Neoplasm: C0006826

Skin cancer: C0007114

USE ONLY THE PATH INFORMATION OBTAINED FROM A TAXONOMY
INFORMATION CONTENT-BASED MEASURES

- Incorporate the probability of the concepts
  - \( IC = -\log(P(\text{concept})) \)
Information Content-Based Measures

- Incorporate the probability of the concepts
  - $IC = -\log(P(\text{concept}))$

- $P(\text{concept})$
  - Calculated by summing the probability of the concept and the probability of its descendants

- Probabilities are obtained from an external corpus
PROBABILITY EXAMPLE

\[ P(\text{Disease [C0012634]}) = \]
\[ P(C0012634) + P(C1302761) + P(C1882062) + P(C0006826) + P(C0007114) + P(C0277579) + P(C0013220); \]

EXTERNAL CORPUS

\[ P(C) = \frac{\text{freq}(C)}{N} \]
INFORMATION CONTENT-BASED MEASURES

- Incorporate the probability of the concepts
  - $IC = -\log(P(\text{concept}))$

- Resnik, 1995
  - $sim(c_1,c_2) = IC(LCS(c_1,c_2))$
INFORMATION CONTENT-BASED MEASURES

- Incorporate the probability of the concepts
  - $IC = -\log(P(\text{concept}))$

- Resnik, 1995
  - $\text{sim}(c1,c2) = IC(\text{LCS}(c2,c2))$

- Jiang and Conrath, 1997
  - $\text{sim}(c1,c2) = \frac{1}{IC(c1) + IC(c2) - 2 \times IC(\text{LCS}(c1,c2))}$
INFORMATION CONTENT-BASED MEASURES

- Incorporate the probability of the concepts
  - $IC = -\log(P(\text{concept}))$

- Resnik, 1995
  - $sim(c_1,c_2) = IC(LCS(c_2,c_2))$

- Jiang and Conrath, 1997
  - $sim(c_1,c_2) = 1 \div (IC(c_1)+IC(c_2) – 2* IC(LCS(c_1,c_2))$

- Lin, 1998
  - $sim(c_1,c_2) = (2*IC(LCS(c_2,c_2))) / (IC(c_1)+IC(c_2))$
SIDE NOTE: COMPARISON BETWEEN LIN AND WU & PALMER

- IC-based measure: Lin, 1998
  \[ \text{sim}(c_1, c_2) = \frac{2 \times \text{IC}(\text{LCS}(c_2, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)} \]

- Path-based measure: Wu and Palmer, 1994
  \[ \text{sim}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_2, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)} \]
SIDE NOTE: COMPARISON BETWEEN LIN AND WU & PALMER

- **IC-based measure: Lin, 1998**
  
  \[ \text{sim}(c_1,c_2) = \frac{(2 \times \text{IC}(\text{LCS}(c_1,c_2)))}{\text{IC}(c_1)+\text{IC}(c_2)} \]

- **Path-based measure: Wu and Palmer, 1994**
  
  \[ \text{sim}(c_1,c_2) = \frac{(2 \times \text{depth}(\text{LCS}(c_1,c_2)))}{\text{depth}(c_1)+\text{depth}(c_2)} \]
IC-BASED SIMILARITY MEASURES

PATH INFORMATION

- Disease: C0012634
- Drug Related Disorder: C0277579
- Drug Tolerance: C0013220
- Neoplasm: C1302761
- Neoplastic Disease: C1882062
- Malignant Neoplasm: C0006826
- Skin cancer: C0007114

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
  - \( \text{rel}(c1, c2) = \sum_{o \in \text{overlap}} \text{length}(o)^2 \)
    - where \( \text{length} = \# \) words in the term

- Contextual information representing its term
  - UMLS Definition
Lesk Measure

- Lesk measure: Lesk, 1986
  - \( \text{rel}(c_1,c_2) = \sum_{o \in \text{overlap}} \text{length}(o)^2 \)
  - where \( \text{length} = \# \text{ words in the term} \)

Finger: C0016129

Any of the terminal digits of the hand

Toe: C0040347

One of the terminal digits of the foot

\( \text{sim}(\text{finger}, \text{toe}) = 2^2 = 4 \)
Vector Measure

- Vector measure: Patwardhan and Pedersen, 2006
  - \( \text{rel}(c_1, c_2) = \cos(2\text{nd-order vector}(c_1), 2\text{nd-order vector}(c_2)) \)
Vector Measure

- Vector measure: Patwardhan and Pedersen, 2006
  - \( \text{rel}(c1, c2) = \cosine(2^{nd}\text{-order vector}(c1), 2^{nd}\text{-order vector}(c2)) \)
VECTOR MEASURE ALGORITHM
VECTOR MEASURE ALGORITHM
Vector Measure Algorithm

One of the five digits of the hand

UMLS

Finger: C0016129
Vector Measure Algorithm

First Order Vectors

One of the five digits of the hand

Finger: C0016129
Vector Measure Algorithm

Ist ORDER VECTORS

ONE OF THE FIVE DIGITS OF THE HAND

Finger: C0016129

2nd ORDER VECTORS

AVERAGE =

UMLS
Vector Measure Algorithm

2nd ORDER VECTORS

One of the five digits of the hand

UMLS

Finger: C0016129
1st ORDER VECTORS

ONE OF THE FIVE DIGITS OF THE HAND

Finger: C0016129
1ST ORDER VECTORS

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

<table>
<thead>
<tr>
<th>Word 1</th>
<th>0</th>
<th>2</th>
<th>0</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 2</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Word 3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Word 4</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Word 5</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Word 6</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Word N</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**One**    **Five**    **Digits**    **Hand**

Finger: C0016129
## 1st Order Vectors

<table>
<thead>
<tr>
<th></th>
<th>ONE</th>
<th>FIVE</th>
<th>DIGITS</th>
<th>HAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Word 2</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Word 3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Word 4</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Word 5</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Word 6</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Word N</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Finger:** C0016129
### 2\textsuperscript{nd} ORDER VECTOR

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 + 2 + 0 + 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 + 5 + 0 + 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 + 0 + 0 + 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 + 0 + 2 + 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 + 0 + 0 + 7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 + 2 + 1 + 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \frac{\text{ONE}}{4} + \frac{\text{FIVE}}{4} + \frac{\text{DIGITS}}{4} + \frac{\text{HAND}}{4} \]

Finger: C0016129
### 2nd Order Vector

<table>
<thead>
<tr>
<th>ONE</th>
<th>FIVE</th>
<th>DIGITS</th>
<th>HAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.25</td>
<td></td>
</tr>
</tbody>
</table>

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

Accuracy

0.55 0.71 0.69 0.7 0.72 0.73 0.74 0.74

baseline path lch wup nam res jcn lin lesk vector

Path-based IC-based Relatedness
COMPARISON ACROSS SUBSETS OF MSH-WSD

Part V

Accuracy

Terms: Baseline 0.55, SenseRelate 0.71, MRD 0.87, 2-MRD 0.85
Acronyms: Baseline 0.54, SenseRelate 0.73, MRD 0.88, 2-MRD 0.93
Mixture: Baseline 0.53, SenseRelate 0.74, MRD 0.8, 2-MRD 0.78
Overall: Baseline 0.55, SenseRelate 0.74, MRD 0.8, 2-MRD 0.78

Legend:
- Baseline
- SenseRelate
- MRD
- 2-MRD
## Comparison Across Subsets of MSH-WSD

<table>
<thead>
<tr>
<th>Subset</th>
<th>Baseline</th>
<th>SenseRelate</th>
<th>MRD</th>
<th>2-MRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms</td>
<td>0.55</td>
<td>0.87</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>Acronyms</td>
<td>0.54</td>
<td>0.87</td>
<td>0.85</td>
<td>0.93</td>
</tr>
<tr>
<td>Mixture</td>
<td>0.53</td>
<td>0.73</td>
<td>0.88</td>
<td>0.78</td>
</tr>
<tr>
<td>Overall</td>
<td>0.55</td>
<td>0.74</td>
<td>0.8</td>
<td>0.78</td>
</tr>
</tbody>
</table>

**Part V**
COMPARISON ACROSS SUBSETS OF MSH-WSD

The diagram compares the accuracy of different methods across subsets of MSH-WSD data. The subsets are:
- Terms
- Acronyms
- Mixture
- Overall

The methods compared are:
- Baseline
- SenseRelate
- MRD
- 2-MRD

The accuracy values are presented for each subset:
- Terms: Baseline 0.55, SenseRelate 0.71, MRD 0.67, 2-MRD 0.71
- Acronyms: Baseline 0.54, SenseRelate 0.73, MRD 0.85, 2-MRD 0.87
- Mixture: Baseline 0.53, SenseRelate 0.88, MRD 0.93, 2-MRD 0.8
- Overall: Baseline 0.55, SenseRelate 0.74, MRD 0.74, 2-MRD 0.78

The diagram visually represents these values, showing the performance of each method across different categories.
COMPARISON ACROSS SUBSETS OF MSH-WSD

Part V
COMPARISON ACROSS SUBSETS OF MSH-WSD

<table>
<thead>
<tr>
<th></th>
<th>Terms</th>
<th>Acronyms</th>
<th>Mixture</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.55</td>
<td>0.54</td>
<td>0.53</td>
<td>0.55</td>
</tr>
<tr>
<td>SenseRelate</td>
<td>0.71</td>
<td>0.85</td>
<td>0.88</td>
<td>0.74</td>
</tr>
<tr>
<td>MRD</td>
<td>0.67</td>
<td>0.87</td>
<td>0.93</td>
<td>0.80</td>
</tr>
<tr>
<td>2-MRD</td>
<td>0.67</td>
<td>0.85</td>
<td>0.88</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Part V
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 accuracy.

![Bar chart showing vector accuracy for different window sizes.](chart.png)
COMPARISON OF WINDOW SIZES FOR LIN

![Bar graph showing the accuracy of different window sizes for LIN. The x-axis represents window sizes ranging from 0 to 70, and the y-axis represents accuracy ranging from 0 to 0.8. The graph shows that the accuracy increases with window size, peaking at approximately 0.74 for window sizes of 25, 50, 60, and 70.]
SURROUNDING TERMS

Not all terms have a concept in the UMLS

d therefore

Not all surrounding terms in the window mapped to CUIs
WINDOW SIZES VERSUS MAPPED TERMS

Number of mappings

Window size

0 0.27 0.79 1.85 3.47 7.57 12.63 14.28 15.56

lin

75
Window sizes versus mapped terms

- Number of mappings
- Window size
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
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

THANK YOU
RESOURCES

- **Software:**
  - UMLS::SenseRelate
    - [http://search.cpan.org/dist/UMLS-SenseRelate/](http://search.cpan.org/dist/UMLS-SenseRelate/)
  - UMLS::Similarity
    - [http://search.cpan.org/dist/UMLS-Similarity/](http://search.cpan.org/dist/UMLS-Similarity/)
  - UMLS::Interface
    - [http://search.cpan.org/dist/UMLS-Interface/](http://search.cpan.org/dist/UMLS-Interface/)

- **Data**
  - MSH-WSD

QUESTIONS?
VECTOR RESULTS WITH DATA SOURCES

Accuracy versus Window Size

NLM Medline Bigrams  UMLS MRCOC  MSH-WSD Medline Data