Lecture Overview

- Semantic Similarity/Distance
- Concepts: Thesaurus/Ontology Methods
- Words: Distributional Methods
Why words/concepts similarity is important?

“fast” is similar to “rapid”
“tall” is similar to “height”

Question answering:

Q: “How tall is Mt. Everest?”
Candidate A: “The official height of Mount Everest is 29029 feet”

• Extends to sentence/paragraph similarity
• **Summarization**: identify and eliminate redundancy, aggregate similar phrase/sentences
• ............
WordNet: entry for “table”

The **noun** "table" has 6 senses in WordNet.

1. table, tabular array (a set of data arranged in rows and columns) "see table 1"
2. table (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs) "it was a sturdy table"
3. table (a piece of furniture with tableware for a meal laid out on it) "I reserved a table at my favorite restaurant"
4. mesa, table (flat tableland with steep edges) "the tribe was relatively safe on the mesa but they had to descend into the valley for water"
5. table (a company of people assembled at a table for a meal or game) "he entertained the whole table with his witty remarks"
6. board, table (food or meals in general) "she sets a fine table"; "room and board"

The **verb** "table" has 1 sense in WordNet.

1. postpone, prorogue, hold over, put over, table, shelve, set back, defer, remit, put off – (hold back to a later time; "let's postpone the exam")
# WordNet Relations (between synsets!)

## Nouns

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyponym</td>
<td>From concepts to subtypes</td>
<td>meal → lunch</td>
</tr>
<tr>
<td>Synonym</td>
<td>From concepts to superordinates</td>
<td>breakfast → meal</td>
</tr>
<tr>
<td>Has-Member</td>
<td>From groups to their members</td>
<td>faculty → professor</td>
</tr>
<tr>
<td>Member-Of</td>
<td>From members to their groups</td>
<td>copilot → crew</td>
</tr>
<tr>
<td>Has-Part</td>
<td>From wholes to parts</td>
<td>table → leg</td>
</tr>
<tr>
<td>Part-Of</td>
<td>From parts to wholes</td>
<td>course → meal</td>
</tr>
<tr>
<td>Antonym</td>
<td>Opposites</td>
<td>leader → follower</td>
</tr>
</tbody>
</table>

## Verbs

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyponym</td>
<td>From events to superordinate events</td>
<td>fly → travel</td>
</tr>
<tr>
<td>Synonym</td>
<td>From events to their subtypes</td>
<td>walk → stroll</td>
</tr>
<tr>
<td>Entails</td>
<td>From events to the events they entail</td>
<td>snore → sleep</td>
</tr>
<tr>
<td>Antonym</td>
<td>Opposites</td>
<td>increase ↔ decrease</td>
</tr>
</tbody>
</table>
Semantic Similarity/Distance: example

(n) table -- (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs)  (n) mesa, table -- (flat tableland with steep edges)

(n) hill (a local and well-defined elevation of the land)  (n) lamp (a piece of furniture holding one or more electric light bulbs)
Semantic Similarity/Distance

Between two concepts in an ontology, e.g., between two senses in Wordnet

What would you use to compute it?

A. The distance between the two concepts in the underlying hierarchies / graphs
B. The glosses of the concepts
C. None of the above
D. Both of the above
Lesk’s (1986) idea: Related word senses are (often) defined *using the same words*. E.g:

- bank(1): “a financial institution”
- bank(2): “sloping land beside a body of water”
- lake: “a body of water surrounded by land”
Gloss Overlaps ≈ Relatedness

Lesk’s (1986) idea: Related word senses are (often) defined *using the same words*. E.g:

- bank(1): “a financial institution”
- bank(2): “sloping land beside a body of water”
- lake: “a body of water surrounded by land”
Gloss Overlaps ≈ Relatedness

Lesk’s (1986) idea: Related word senses are (often) defined *using the same words*. E.g:

- bank(1): “a financial institution”
- bank(2): “sloping land beside a body of water”
- lake: “a body of water surrounded by land”

Gloss overlaps = # content words common to two glosses ≈ relatedness

- Thus, relatedness (bank(2), lake) = 3
- And, relatedness (bank(1), lake) = 0
Limitations of (Lesk’s) Gloss Overlaps

► Most glosses are very short.
  ▪ So not enough words to find overlaps with.

► Solution?
Extended gloss overlaps
  ▪ Add glosses of synsets connected to the input synsets.
Extending a Gloss

**sentence**: “the penalty meted out to one adjudged guilty”

**bench**: “persons who hear cases in a court of law”

# overlapped words = 0
Extending a Gloss

**final judgment**: “a judgment disposing of the case before the court of law”

**sentence**: “the penalty meted out to one adjudged guilty”

**bench**: “persons who hear cases in a court of law”

# overlapped words = 0
Extending a Gloss

final judgment: “a judgment disposing of the case before the court of law”

sentence: “the penalty meted out to one adjudged guilty”

bench: “persons who hear cases in a court of law”

# overlapped words = 2
Creating the Extended Gloss Overlap Measure

► How to measure overlaps?

► Which relations to use for gloss extension?
How to Score Overlaps?

► Lesk simply summed up overlapped words.
► But matches involving phrases – phrasal matches – are rarer, and more informative
  ▪ E.g. “court of law” “body of water”
► Aim: Score of $n$ words in a phrase $> \sum$ of scores of $n$ words in shorter phrases
► Solution: Give a phrase of $n$ words a score of $n^2$
  ▪ “court of law” gets score of 9.
  ▪ bank(2): “sloping land beside a body of water”
  ▪ lake: “a body of water surrounded by land”
Which Relations to Use?

Typically include...

► Hyponyms [ “car” → “vehicle” ]
► Hyponyms [ “car” → “convertible” ]
► Meronyms [ “car” → “accelerator” ]
► …
Extended Gloss Overlap Measure

- Input two synsets A and B
- Find phrasal gloss overlaps between A and B
- For each relation, compute phrasal gloss overlaps between every synset connected to A, and every synset connected to B

A and B can be from different parts of speech!
Distance: Path-length
Path-length sim based on is-a/hypernyms hierarchies

$$\text{sim}_{\text{path}}(c_1, c_2) = 1 / \text{pathlen}(c_1, c_2)$$

$c_1, c_2$ are senses

But this is assuming that all the links are the same…. Encode the same semantic distance…. 
Probability of a concept/sense and its info content

\[ \text{IC}(c) = - \log P(c) \]

\[ P(c) = \frac{\text{count}(c)}{N} \]

\[ \text{IC}(\text{thing}) = 0 \]

\[ \text{IC}(\text{snake}) = 23 \]
Concept Distance: info content

- Similarity should be proportional to the information that the two concepts share... what is that?

\[ P(\text{root}) = 1 \]

The lower the concept/sense, the lower its probability.

\[ P(c) = \frac{\text{count}(c)}{N} \]

\[ \text{IC}(c) = -\log P(c) \]

\[ \text{IC}(c) = -\log P(\text{LCS}(c_1, c_2)) \]

\[ \text{LCS}(c_1, c_2) \]

\[ \text{probability} \]

Information Content

Lowest Common Subsumer

\[ \text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2)) \]
Given this measure of similarity

\[ \text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(LCS(c_1, c_2)) \]

Are these two the same?

- A. Yes
- B. No
- C. Cannot tell

Is this reasonable?

- A. Yes
- B. No
- C. Cannot tell
Concept Distance: info content

- One of best performers - Jiang-Conrath distance
- How much **information the two DO NOT share**?

![Diagram showing concept distance calculation]

\[
IC(c_1) + IC(c_2) - 2 \times IS(c_1, c_2) + (IC(c_1) - IS(c_1, c_2)) + (IC(c_2) - IS(c_1, c_2))
\]
Concept Distance: info content

\[
\begin{align*}
(\text{IC}(c_1) - \text{IS}(c_1,c_2)) + (\text{IC}(c_2) - \text{IS}(c_1,c_2)) \\
\sqrt{\text{IC}(c_1) + \text{IC}(c_2) - 2 \times \text{IS}(c_1,c_2)}
\end{align*}
\]

\[
\text{dist}_{JC}(c_1, c_2) = (\log P(c_1)) + (\log P(c_2)) + (2 \times \log P(\text{LCS}(c_1, c_2)))
\]

\[
\text{dist}_{JC}(c_1, c_2) = 2 \times \log P(\text{LCS}(c_1, c_2)) - (\log P(c_1) + \log P(c_2))
\]

- This is a measure of distance. Reciprocal for similarity.
- Problem for measures working on hierarchies/graphs: only compare concepts associated with words of one part-of-speech (typically nouns)
Concept Distance: info content

- One of best performers - Jiang-Conrath distance
- How much information the two DO NOT share

$$\text{dist}_{JC}(c_1, c_2) = (-(\log P(c_1)) + (-\log P(c_2))) - (2 \times -\log P(\text{LCS}(c_1, c_2)))$$

$$\text{dist}_{JC}(c_1, c_2) = 2 \times \log P(\text{LCS}(c_1, c_2)) - (\log P(c_1) + \log P(c_2))$$

- This is a measure of distance. Reciprocal for similarity! $\frac{1}{\text{dist}_{JC}}$
- Problem for measures working on hierarchies/graphs: only compare concepts associated with words of one part-of-speech (typically nouns)
Example

\[
\text{sim}_{\text{res}}(\text{Dog, Snake}) = 13
\]

\[
\text{sim}_{\text{res}}(\text{Mammal, Reptile}) = 13
\]

\[
\text{dist}_{\text{JC}}(\text{Dog, Snake}) = (2\times 13) + (22 + 23) = 19
\]

\[
\text{dist}_{\text{JC}}(\text{Mammal, Reptile}) = (2\times 13) + (19 + 20) = 13
\]
Lecture Overview

- Semantic Similarity/Distance
- Concepts: Thesaurus/Ontology Methods
- Words: Distributional Methods – Word Similarity (WS)
Word Similarity: Distributional Methods

• Do not have any thesauri/ontologies for target language (e.g., Russian)

• If you have thesaurus/ontology, still
  - Missing domain-specific (e.g., technical words)
  - Poor hyponym knowledge (for V) and nothing for Adj and Adv
  - Difficult to compare senses from different hierarchies (although extended Lesk can do this)

• Solution: extract similarity from corpora

• Basic idea: two words are similar if they appear in similar contexts
Intuition of distributional word similarity

Example: Suppose I asked you what is tesgüino?

- A bottle of tesgüino is on the table
- Everybody likes tesgüino
- Tesgüino makes you drunk
- We make tesgüino out of corn.

From context words humans can guess tesgüino means
- an alcoholic beverage like beer

Intuition for algorithm:
- Two words are similar if they have similar word contexts.
Word-Word matrix: Sample contexts ± 7 words

sugar, a sliced lemon, a tablespoonful of their enjoyment. Cautiously she sampled her first well suited to programming on the digital for the purpose of gathering data and preserve or jam, a pinch each of, and another fruit whose taste she likened In finding the optimal R-stage policy from necessary for the study authorized in the

• Portion of matrix from the Brown corpus

<table>
<thead>
<tr>
<th>aardvark</th>
<th>computer</th>
<th>data</th>
<th>pinch</th>
<th>result</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>apricot</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>pineapple</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>digital</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>information</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Simple example of Vectors Models aka “embeddings”.

• Model the meaning of a word by “embedding” in a vector space.
• The meaning of a word is a vector of numbers
WS Distributional Methods (2)

- More informative values (referred to as weights or measure of association in the literature)
- Point-wise Mutual Information

\[
assoc_{PMI} (w, w_i) = \log_2 \frac{P(w, w_i)}{P(w)P(w_i)}
\]

- \(t\)-test

\[
assoc_t \text{-test} (w, w_i) = \frac{P(w, w_i) - P(w)P(w_i)}{\sqrt{P(w)P(w_i)}}
\]
Positive Pointwise Mutual Information

- **PMI** ranges from \(-\infty\) to \(+\infty\)
- But the negative values are problematic
  - Things are co-occurring less than we expect by chance
  - Unreliable without enormous corpora
    - Imagine \(w_1\) and \(w_2\) whose probability is each \(10^{-6}\)
    - Hard to be sure \(p(w_1, w_2)\) is significantly different than \(10^{-12}\)
  - Plus it's not clear people are good at "unrelatedness"
- So we just replace negative PMI values by 0
- Positive PMI (PPMI) between \(word_1\) and \(word_2\):

\[
PPMI(word_1, word_2) = \max \left( \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}, 0 \right)
\]
PMI example

\[ \text{assoc}_{\text{PMI}}(w, w_i) = \log_2 \frac{P(w, w_i)}{P(w)P(w_i)} \]

Assume \( w, w_i \) appear with equal frequency \( \frac{1}{2^{10}} \) in a large set of documents.

\[
\begin{align*}
P(w) &= 2^{-10} \\
P(w_i) &= 2^{-10} \\
P(w, w_i) &= 2^{-10} \cdot 2^{-10} = 2^{-20} \quad \text{if the words are completely independent} \\
\end{align*}
\]

\[
\begin{align*}
P(w, w_i) &= 2^{-10} \quad \text{if the words appear always together} \\
\end{align*}
\]

\[
\begin{align*}
A \quad \text{assoc}_{\text{PMI}} &= \log_2 \frac{2^{-20}}{2^{-10} \cdot 2^{-10}} = \log_2 1 = 0 \\
B \quad \text{assoc}_{\text{PMI}} &= \log_2 \frac{2^{-10}}{2^{-10} \cdot 2^{-10}} = \log_2 2^{10} = 10 \\
\end{align*}
\]
Other popular vector representations

Dense vector representations (less dimensions):

1. Singular value decomposition applied to word-word PointWise-MI matrix
2. Neural-Network-inspired models (skip-grams, CBOW)
WS Distributional Methods (3)

- Similarity between vectors

\[
sim_{\text{cosine}}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{||\vec{v}|| \cdot ||\vec{w}||} = \frac{\vec{v} \cdot \vec{w}}{||\vec{v}|| \times ||\vec{w}||} = \cos(\alpha)
\]

Not sensitive to extreme values

\[
sim_{\text{Jaccard}}(\vec{v}, \vec{w}) = \frac{\sum_{i=1}^{N} \min(v_i, w_i)}{\sum_{i=1}^{N} \max(v_i, w_i)}
\]

Normalized (weighted) number of overlapping features

\[
\frac{2 + 1 + 0 + 2}{3 + 1 + 1 + 3} = \frac{5}{8}
\]

\[
\frac{3 + 1 + 1}{3} = \frac{5}{3}
\]
Learning Goals for today’s class

You can:

• Describe and Justify metrics to compute the similarity/distance of two concepts in an ontology

• Describe and Justify distributional metrics to compute the similarity/distance of two words (or phrases) in a Natural Language
Assignment-3 out - due Mar 30
(8-18 hours - working in pairs on programming parts is strongly advised)

Midterm marking almost done! Likely completed by tomorrow

Next class Wed

- Natural language Processing (Context free grammars and parsing)