Word Meaning and Similarity

Word Senses and Word Relations

Slides are adapted from Dan Jurafsky
Reminder: lemma and wordform

• A **lemma** or **citation form**
  • Same stem, part of speech, rough semantics

• A **wordform**
  • The “inflected” word as it appears in text

<table>
<thead>
<tr>
<th>Wordform</th>
<th>Lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>banks</td>
<td>bank</td>
</tr>
<tr>
<td>sung</td>
<td>sing</td>
</tr>
<tr>
<td>duermes</td>
<td>dormir</td>
</tr>
</tbody>
</table>
Lemmas have senses

• One lemma “bank” can have many meanings:

  Sense 1:  • …a *bank* can hold the investments in a custodial account…
  Sense 2:  • “…as agriculture burgeons on the east *bank* the river will shrink even more”

• Sense (or word sense)
  • A discrete representation of an aspect of a word’s meaning.

• The lemma *bank* here has two senses
Homonymy

Homonyms: words that share a form but have unrelated, distinct meanings:

- \(\text{bank}_1\): financial institution, \(\text{bank}_2\): sloping land
- \(\text{bat}_1\): club for hitting a ball, \(\text{bat}_2\): nocturnal flying mammal

1. Homographs (bank/bank, bat/bat)
2. Homophones:
   1. Write and right
   2. Piece and peace
Homonymy causes problems for NLP applications

- Information retrieval
  - “bat care”
- Machine Translation
  - bat: *murciélago* (animal) or *bate* (for baseball)
- Text-to-Speech
  - bass (stringed instrument) vs. *bass* (fish)
Polysemy

1. The **bank** was constructed in 1875 out of local red brick.
2. I withdrew the money from the **bank**

Are those the same sense?
- Sense 2: “A financial institution”
- Sense 1: “The building belonging to a financial institution”

A **polysemous** word has **related** meanings
- Most non-rare words have multiple meanings
Metonymy or Systematic Polysemy: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - School, university, hospital
  - All can mean the institution or the building.

- A systematic relationship:
  - Building ↔ Organization

- Other such kinds of systematic polysemy:
  - Author (Jane Austen wrote Emma)
    ↔ Works of Author (I love Jane Austen)
  - Tree (Plums have beautiful blossoms)
    ↔ Fruit (I ate a preserved plum)
How do we know when a word has more than one sense?

• The “zeugma” test: Two senses of serve?
  • Which flights serve breakfast?
  • Does Lufthansa serve Philadelphia?
  • Does Lufthansa serve breakfast and San Jose?

• Since this conjunction sounds weird,
  • we say that these are two different senses of “serve”
Synonyms

- Word that have the same meaning in some or all contexts.
  - filbert / hazelnut
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - Water / $H_2O$

- Two lexemes are synonyms
  - if they can be substituted for each other in all situations
  - If so they have the same *propositional meaning*
**Synonyms**

- But there are few (or no) examples of perfect synonymy.
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.

- Example:
  - Water/H₂O
  - Big/large
  - Brave/courageous
Synonymy is a relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How *big* is that plane?
  - Would I be flying on a *large* or small plane?
- How about here:
  - Miss Nelson became a kind of *big* sister to Benjamin.
  - Miss Nelson became a kind of *large* sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense
Antonyms

• Senses that are opposites with respect to one feature of meaning
• Otherwise, they are very similar!
  dark/light  short/long  fast/slow  rise/fall
  hot/cold    up/down    in/out
• More formally: antonyms can
  • define a binary opposition
    or be at opposite ends of a scale
    • long/short, fast/slow
• Be reversives:
  • rise/fall, up/down
Hyponymy and Hypernymy

- One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other
  - *car* is a hyponym of *vehicle*
  - *mango* is a hyponym of *fruit*
- Conversely **hypernym/superordinate** ("hyper is super")
  - *vehicle* is a hypernym of *car*
  - *fruit* is a hypernym of *mango*

<table>
<thead>
<tr>
<th>Superordinate/hyper</th>
<th>vehicle</th>
<th>fruit</th>
<th>furniture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subordinate/hyponym</td>
<td>car</td>
<td>mango</td>
<td>chair</td>
</tr>
</tbody>
</table>
Hyponymy more formally

- Extensional:
  - The class denoted by the superordinate extensionally includes the class denoted by the hyponym

- Entailment:
  - A sense A is a hyponym of sense B if being an A entails being a B

- Hyponymy is usually transitive
  - (A hypo B and B hypo C entails A hypo C)

- Another name: the IS-A hierarchy
  - A IS-A B (or A ISA B)
  - B subsumes A
Hyponyms and Instances

- WordNet has both **classes** and **instances**.
- An **instance** is an individual, a proper noun that is a unique entity
  - *San Francisco* is an **instance** of *city*
- But *city* is a class
  - *city* is a **hyponym** of *municipality*...*location*...
Word Meaning and Similarity

Word Senses and Word Relations
Word Meaning and Similarity

WordNet
Applications of Thesauri and Ontologies

• Information Extraction
• Information Retrieval
• Question Answering
• Bioinformatics and Medical Informatics
• Machine Translation
WordNet 3.0

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
  - Some other languages available or under development
    - (Arabic, Finnish, German, Portuguese...)

<table>
<thead>
<tr>
<th>Category</th>
<th>Unique Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117,798</td>
</tr>
<tr>
<td>Verb</td>
<td>11,529</td>
</tr>
<tr>
<td>Adjective</td>
<td>22,479</td>
</tr>
<tr>
<td>Adverb</td>
<td>4,481</td>
</tr>
</tbody>
</table>
Senses of “bass” in Wordnet

Noun

- **S: (n) bass** (the lowest part of the musical range)
- **S: (n) bass, bass part** (the lowest part in polyphonic music)
- **S: (n) bass, basso** (an adult male singer with the lowest voice)
- **S: (n) sea bass, bass** (the lean flesh of a saltwater fish of the family Serranidae)
- **S: (n) freshwater bass, bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- **S: (n) bass, bass voice, basso** (the lowest adult male singing voice)
- **S: (n) bass** (the member with the lowest range of a family of musical instruments)
- **S: (n) bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- **S: (adj) bass, deep** (having or denoting a low vocal or instrumental range) "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"
How is “sense” defined in WordNet?

• The synset (synonym set), the set of near-synonyms, instantiates a sense or concept, with a gloss

• Example: chump as a noun with the gloss:
  “a person who is gullible and easy to take advantage of”

• This sense of “chump” is shared by 9 words:
  chump¹, fool², gull¹, mark⁹, patsy¹, fall guy¹, sucker¹, soft touch¹, mug²

• Each of these senses have this same gloss
  • (Not every sense; sense 2 of gull is the aquatic bird)
WordNet Hypernym Hierarchy for “bass”

- S: (n) bass, basso (an adult male singer with the lowest voice)
  - direct hypernym / inherited hypernym / sister term
    - S: (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
      - S: (n) musician, instrumentalist, player (someone who plays a musical instrument (as a profession))
      - S: (n) performer, performing artist (an entertainer who performs a dramatic or musical work for an audience)
        - S: (n) entertainer (a person who tries to please or amuse)
      - S: (n) person, individual, someone, somebody, mortal, soul (a human being) "there was too much for one person to do"
    - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
      - S: (n) living thing, animate thing (a living (or once living) entity)
        - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?"; "the team is a unit"
      - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
        - S: (n) physical entity (an entity that has physical existence)
      - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
# WordNet Noun Relations

<table>
<thead>
<tr>
<th>Relation</th>
<th>Also called</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyponym</td>
<td>Subordinate</td>
<td>From concepts to subtypes</td>
<td>meal¹ → lunch¹</td>
</tr>
<tr>
<td>Member Meronym</td>
<td>Has-Member</td>
<td>From groups to their members</td>
<td>faculty² → professor¹</td>
</tr>
<tr>
<td>Has-Instance</td>
<td></td>
<td>From concepts to instances of the concept</td>
<td>composer¹ → Bach¹</td>
</tr>
<tr>
<td>Instance</td>
<td></td>
<td>From instances to their concepts</td>
<td>Austen¹ → author¹</td>
</tr>
<tr>
<td>Member Holonym</td>
<td>Member-Of</td>
<td>From members to their groups</td>
<td>copilot¹ → crew¹</td>
</tr>
<tr>
<td>Part Meronym</td>
<td>Has-Part</td>
<td>From wholes to parts</td>
<td>table² → leg³</td>
</tr>
<tr>
<td>Part Holonym</td>
<td>Part-Of</td>
<td>From parts to wholes</td>
<td>course⁷ → meal¹</td>
</tr>
<tr>
<td>Antonym</td>
<td></td>
<td>Opposites</td>
<td>leader¹ → follower¹</td>
</tr>
</tbody>
</table>
WordNet 3.0

• Where it is:
  • http://wordnetweb.princeton.edu/perl/webwn

• Libraries
  • Python: WordNet from NLTK
    • http://www.nltk.org/Home
  • Java:
    • JWNL, extJWNL on sourceforge
Word Meaning and Similarity

WordNet
Word Meaning and Similarity

Word Similarity: Thesaurus Methods
Word Similarity

- **Synonymy**: a binary relation
  - Two words are either synonymous or not
- **Similarity** (or distance): a looser metric
  - Two words are more similar if they share more features of meaning
- Similarity is properly a relation between *senses*
  - The word “bank” is not similar to the word “slope”
  - Bank$^1$ is similar to fund$^3$
  - Bank$^2$ is similar to slope$^5$
- But we’ll compute similarity over both words and senses
Why word similarity

- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading
- Plagiarism detection
- Document clustering
Word similarity and word relatedness

• We often distinguish word similarity from word relatedness
  • Similar words: near-synonyms
  • Related words: can be related any way
    • car, bicycle: similar
    • car, gasoline: related, not similar
Two classes of similarity algorithms

• Thesaurus-based algorithms
  • Are words “nearby” in hypernym hierarchy?
  • Do words have similar glosses (definitions)?

• Distributional algorithms
  • Do words have similar distributional contexts?
Path based similarity

• Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy
  • =have a short path between them
  • concepts have path 1 to themselves
Refinements to path-based similarity

- $\text{pathlen}(c_1, c_2) = 1 + \text{number of edges in the shortest path in the hypernym graph between sense nodes } c_1 \text{ and } c_2$
- ranges from 0 to 1 (identity)

- $\text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)}$

- $\text{wordsim}(w_1, w_2) = \max_{c_1 \in \text{senses}(w_1), c_2 \in \text{senses}(w_2)} \text{simpath}(c_1, c_2)$
Example: path-based similarity

$$\text{simpath}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)}$$

$$\text{simpath}(\text{nickel}, \text{coin}) = \frac{1}{2} = .5$$

$$\text{simpath}(\text{fund}, \text{budget}) = \frac{1}{2} = .5$$

$$\text{simpath}(\text{nickel}, \text{currency}) = \frac{1}{4} = .25$$

$$\text{simpath}(\text{nickel}, \text{money}) = \frac{1}{6} = .17$$

$$\text{simpath}(\text{coinage}, \text{Richter scale}) = \frac{1}{6} = .17$$
Problem with basic path-based similarity

• Assumes each link represents a uniform distance
  • But *nickel* to *money* seems to us to be closer than *nickel* to *standard*
  • Nodes high in the hierarchy are very abstract

• We instead want a metric that
  • Represents the cost of each edge independently
  • Words connected only through abstract nodes
    • are less similar
Information content similarity

- Train by counting in a corpus
  - Each instance of hill counts toward frequency of natural elevation, geological formation, entity, etc
  - Let words(c) be the set of all words that are children of node c
    - words("geo-formation") = {hill,ridge,grotto,coast,cave,shore,natural elevation}
    - words("natural elevation") = {hill, ridge}

\[
P(c) = \frac{\sum_{w \in \text{words}(c)} \text{count}(w)}{N}
\]
Information content similarity

• WordNet hierarchy augmented with probabilities $P(c)$

Information content: definitions

- Information content:
  \[ IC(c) = -\log P(c) \]

- Most informative subsumer (Lowest common subsumer)
  \[ LCS(c_1, c_2) = \]
  The most informative (lowest) node in the hierarchy subsuming both \( c_1 \) and \( c_2 \)
Using information content for similarity: the Resnik method


• The similarity between two words is related to their common information
• The more two words have in common, the more similar they are
• Resnik: measure common information as:
  • The information content of the most informative (lowest) subsumer (MIS/LCS) of the two nodes
  • $\text{sim}_{\text{resnik}}(c_1,c_2) = -\log P(\text{LCS}(c_1,c_2))$
Dekang Lin method

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. ICML

- Intuition: Similarity between A and B is not just what they have in common
- The more **differences** between A and B, the less similar they are:
  - Commonality: the more A and B have in common, the more similar they are
  - Difference: the more differences between A and B, the less similar
- Commonality: $IC(\text{common}(A,B))$
- Difference: $IC(\text{description}(A,B)) - IC(\text{common}(A,B))$
Dekang Lin similarity theorem

- The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are:

\[
\text{sim}_{Lin}(A, B) \propto \frac{IC(\text{common}(A, B))}{IC(\text{description}(A, B))}
\]

- Lin (altering Resnik) defines IC(common(A,B)) as 2 x information of the LCS:

\[
\text{sim}_{Lin}(c_1, c_2) = \frac{2 \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)}
\]
Lin similarity function

\[ sim_{Lin}(A, B) = \frac{2 \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \]

\[ sim_{Lin}(\text{hill, coast}) = \frac{2 \log P(\text{geological-formation})}{\log P(\text{hill}) + \log P(\text{coast})} \]

\[ = \frac{2 \ln 0.00176}{\ln 0.0000189 + \ln 0.0000216} \]

\[ = .59 \]
The (extended) Lesk Algorithm

- A thesaurus-based measure that looks at glosses
- Two concepts are similar if their glosses contain similar words
  - *Drawing paper*: paper that is *specially prepared* for use in drafting
  - *Decal*: the art of transferring designs from *specially prepared paper* to a wood or glass or metal surface
- For each $n$-word phrase that’s in both glosses
  - Add a score of $n^2$
  - *Paper* and *specially prepared* for $1 + 2^2 = 5$
  - Compute overlap also for other relations
    - glosses of hypernyms and hyponyms
Summary: thesaurus-based similarity

\[ \text{sim}_{\text{path}}(c_1, c_2) = \frac{1}{\text{pathlen}(c_1, c_2)} \]

\[ \text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(\text{LCS}(c_1, c_2)) \quad \text{sim}_{\text{lin}}(c_1, c_2) = \frac{2 \log P(\text{LCS}(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \]

\[ \text{sim}_{\text{eLesk}}(c_1, c_2) = \sum_{r, q \in \text{RELS}} \text{overlap} \left( \text{gloss}(r(c_1)), \text{gloss}(q(c_2)) \right) \]
Libraries for computing thesaurus-based similarity

- **NLTK**

- **WordNet::Similarity**
  - Web-based interface:
    - [http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi](http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi)
Evaluating similarity

- **Extrinsic (task-based, end-to-end) Evaluation:**
  - Question Answering
  - Spell Checking
  - Essay grading

- **Intrinsic Evaluation:**
  - Correlation between algorithm and human word similarity ratings
    - Wordsim353: 353 noun pairs rated 0-10. \( sim(\text{plane}, \text{car})=5.77 \)
  - Taking TOEFL multiple-choice vocabulary tests
    - *Levied* is closest in meaning to:
      - imposed, believed, requested, correlated
Word Meaning and Similarity

Word Similarity: Thesaurus Methods