NLP
Text Similarity

Wordnet
Wordnet

• Wordnet is a project run by George Miller (1920–2012) and Christiane Fellbaum at Princeton University.
• It includes a database of words (mainly nouns and verbs but also adjectives and adverbs) and semantic relations between them.
• The main relation is hypernymy, so the overall structure of the database is more tree-like (see next slide).
• References:
Tree-like structure of Wordnet

- **Ungulate**
  - **Even-toed Ungulate**
    - **Ruminant**
      - Okapi
      - Deer
      - Giraffe
      - Elk
      - Wapiti
      - Caribou
  - **Odd-toed Ungulate**
    - **Equine**
      - Mule
      - Horse
      - Zebra
      - Pony
The noun bar has 11 senses

1. barroom, bar, saloon, ginmill, taproom -- (a room where alcoholic drinks are served over a counter)
2. bar -- (a counter where you can purchase food or drink)
3. bar -- (a rigid piece of metal)
4. measure, bar -- (notation for a repeating pattern of musical beats; written followed by a vertical bar)
5. bar -- (usually metal placed in windows to prevent escape)
6. prevention, bar -- (the act of preventing)
7. bar -- (a unit of pressure equal to a million dynes per square centimeter)
8. bar -- (a submerged (or partly submerged) ridge in a river or along a shore)
9. legal profession, bar, legal community -- (the body of individuals qualified to practice law)
10. cake, bar -- (a block of soap or wax)
11. bar -- ((law) a railing that encloses the part of the courtroom where the judges and lawyers sit and the case is tried)

The verb bar has 4 senses

1. bar, debar, exclude -- (prevent from entering; keep out; "He was barred from membership in the club")
2. barricade, block, blockade, block off, block up, bar -- (render unsuitable for passage; "block the way"; "barricade the streets")
3. banish, relegate, bar -- (expel, as if by official decree; "he was banished from his own country")
4. bar -- (secure with, or as if with, bars; "He barred the door")
Sense 1
barroom, bar, saloon, ginmill, taproom
  => room
  => area
  => structure, construction
    => artifact, artefact
    => object, physical object
      => entity, something

Sense 2
bar
  => counter
  => table
    => furniture, piece of furniture, article of furniture
      => furnishings
        => instrumentality, instrumentation
          => artifact, artefact
            => object, physical object
              => entity, something
Wordnet Example (3/6)

Sense 3
bar
  => implement
     => instrumentality, instrumentation
     => artifact, artefact
     => object, physical object
     => entity, something

Sense 4
measure, bar
  => musical notation
     => notation, notational system
     => writing, symbolic representation
     => written communication, written language
     => communication
     => social relation
     => relation
     => abstraction
Wordnet Example (4/6)

Sense 5
bar
  => obstruction, impediment, impedimenta
  => structure, construction
    => artifact, artefact
      => object, physical object
        => entity, something

Sense 6
prevention, bar
  => hindrance, interference, interfering
    => act, human action, human activity

Sense 7
bar
  => pressure unit
    => unit of measurement, unit
      => definite quantity
        => measure, quantity, amount, quantum
          => abstraction
Wordnet Example (5/6)

Sense 8
bar
    => ridge
         => natural elevation, elevation
         => geological formation, geology, formation
             => natural object
                 => object, physical object
                     => entity, something
    => barrier
        => mechanism
        => natural object
            => object, physical object
                => entity, something
Wordnet Example (6/6)

Sense 9
legal profession, bar, legal community
  => profession, community
  => occupation, vocation, occupational group
    => body
    => gathering, assemblage
    => social group
    => group, grouping

Sense 10
cake, bar
  => block
    => artifact, artefact
    => object, physical object
    => entity, something
Familiarity and Polysemy

board used as a noun is familiar (polysemy count = 9)

bird used as a noun is common (polysemy count = 5)

cat used as a noun is common (polysemy count = 7)

house used as a noun is familiar (polysemy count = 11)

information used as a noun is common (polysemy count = 5)

retrieval used as a noun is uncommon (polysemy count = 3)

serendipity used as a noun is very rare (polysemy count = 1)
## Top-Level Categories

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP</td>
<td>STATIVE</td>
</tr>
<tr>
<td>PERSON</td>
<td>COGNITION</td>
</tr>
<tr>
<td>ARTIFACT</td>
<td>COMMUNIC.*</td>
</tr>
<tr>
<td>COGNITION</td>
<td>SOCIAL</td>
</tr>
<tr>
<td>FOOD</td>
<td>MOTION</td>
</tr>
<tr>
<td>ACT</td>
<td>POSSESSION</td>
</tr>
<tr>
<td>LOCATION</td>
<td>CHANGE</td>
</tr>
<tr>
<td>TIME</td>
<td>EMOTION</td>
</tr>
<tr>
<td>EVENT</td>
<td>PERCEPTION</td>
</tr>
<tr>
<td>COMMUNIC.*</td>
<td>CONSUMPTION</td>
</tr>
<tr>
<td>POSSESSION</td>
<td>BODY</td>
</tr>
<tr>
<td>ATTRIBUTE</td>
<td>TRANSACTION</td>
</tr>
<tr>
<td>QUANTITY</td>
<td>COMPETITION</td>
</tr>
<tr>
<td>ANIMAL</td>
<td>CONTACT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>place</td>
<td>hair</td>
</tr>
<tr>
<td>people</td>
<td>pain</td>
</tr>
<tr>
<td>car</td>
<td>flower</td>
</tr>
<tr>
<td>way</td>
<td>portion</td>
</tr>
<tr>
<td>food</td>
<td>oil</td>
</tr>
<tr>
<td>service</td>
<td>discomfort</td>
</tr>
<tr>
<td>area</td>
<td>process</td>
</tr>
<tr>
<td>day</td>
<td>reason</td>
</tr>
<tr>
<td>experience</td>
<td>result</td>
</tr>
<tr>
<td>review</td>
<td>get... done</td>
</tr>
</tbody>
</table>

[Example from J&M, based on Schneider and Smith 2013]
Text Similarity

Other Lexical Networks
Medical Subject Headings

External links

• EuroWordNet

• Open Thesaurus
  – http://www.openthesaurus.de/

• Freebase
  – http://www.freebase.com

• DBPedia
  – http://www.dbpedia.org

• BabelNet
  – http://babelnet.org

• Various thesauri
  – https://sites.google.com/site/openrogets/
NLP