Agenda

- Discussion of HW5 (due date changed to Thursday)
- Questions, comments, concerns?
- Finish semantics discussion
- Text normalization

Intuition of Semantic Similarity

<table>
<thead>
<tr>
<th>Semantically close</th>
<th>Semantically distant</th>
</tr>
</thead>
<tbody>
<tr>
<td>bank–money</td>
<td>doctor–beer</td>
</tr>
<tr>
<td>apple–fruit</td>
<td>painting–January</td>
</tr>
<tr>
<td>tree–forest</td>
<td>money–river</td>
</tr>
<tr>
<td>bank–river</td>
<td>apple–penguin</td>
</tr>
<tr>
<td>pen–paper</td>
<td>nurse–bottle</td>
</tr>
<tr>
<td>run–walk</td>
<td>pen–river</td>
</tr>
<tr>
<td>mistake–error</td>
<td>clown–tramway</td>
</tr>
<tr>
<td>car–wheel</td>
<td>car–algebra</td>
</tr>
</tbody>
</table>

Why?

- Meaning
  - The two concepts are close in terms of their meaning
- World knowledge
  - The two concepts have similar properties, often occur together, or occur in similar contexts
- Psychology
  - We often think of the two concepts together

Two Types of Relations

- Synonymy: two words are (roughly) interchangeable
- Semantic similarity (distance): somehow "related"
  - Sometimes, explicit lexical semantic relationship, often, not

Validity of Semantic Similarity

- Is semantic distance a valid linguistic phenomenon?
- Experiment (Rubenstein and Goodenough, 1965)
  - Compiled a list of word pairs
  - Subjects asked to judge semantic distance (from 0 to 4) for each of the word pairs
- Results:
  - Rank correlation between subjects is ~0.9
  - People are consistent!
**Compute Semantic Similarity?**

- Task: automatically compute semantic similarity between words
- Theoretically useful for many applications:
  - Detecting paraphrases (i.e., automatic essay grading, plagiarism detection)
  - Information retrieval
  - Machine translation
  - ...
  - Solution in search of a problem?

**Agenda**

- Discussion of HW5 (due date changed to Thursday)
- Questions, comments, concerns?
- Finish semantics discussion
- Text normalization
  - Tokenization
  - Abbreviations
  - Misspellings

**Handling Real-World Text**

- Real-world text is messy
- Often much pre-processing (and post-processing) performed to make real-world text easier to handle
  - Dirty little secret of NLP
  - Not typically viewed as a research area

**Handling Real-World Text**

- Real text is messy
  - Real estate ads:
    50's Sutton Place Area Convertible 3BR 1400 SF 2BR, 2Bth, L-Shaped LR, S.E. Open Vus, Gar, Rf Dk, Mid $400K's Thompson Kane Ina 339-8300

**Tokenization**

- Baseline strategy in English: white space delimited text
  - Not uniformly successful: special characters and punctuation
  - Not applicable in other writing systems, e.g., Chinese
- Question: what is the token identity?
  - Some orthographic information is lexical, some not
  - Include token-initial or token-final punctuation? (Abbrev.)
  - How about capitalization information? (The vs. Mr.)
  - Even without fixed lexicon, want multiple instances of the same term to be recognized, including abbreviations

**(Pre-)Processing Raw Text**

- Amount of pre-processing will differ depending on domain
  - Free text clinical notes may be very messy
  - Journal articles may have significant markup
- First requirement of any approach: **tokenization**
  - Find word boundaries, establish base "tokens" in strings
  - Given tokens, further processing can be done
    - Mapping tokens to terms in a lexicon
    - Word sense disambiguation
    - Extraction of multi-token terms
Lexicon

- Lexicon can be explicit or implicit (or both)
  - Implicit lexicon instantiated on the fly, e.g., $10,205,417
  - May want to assign classes to tokens (e.g., 'Number')
- Explicit lexicon may be built in various ways
  - Manually curated by experts (e.g., WordNet)
  - From large corpora (data driven)
- Given a large explicit lexicon, failure to find exact match:
  - Some text normalization required
  - Belongs to large set deliberately excluded (e.g., numbers)
  - Genuine out-of-vocabulary (OOV) item (e.g., proper name)

Domain-Specific Lexicons

- If the task is bounded in some way, this is helpful
- A key method for bounding processing is domain specification
- If the domain is known (e.g., movies) can exploit narrow knowledge bases more effectively
  - 'Travolta' and 'gaffer' should be in lexicon
  - Not so interested in 'Kahlenberger' or 'perceptron'
- For areas where large curated resources exist, can be very helpful
  - In particular, named entity extraction can benefit

Text Normalization

- Transform a text sequence to improve overall consistency
  - Capitalization normalization
  - Removal of punctuation and other extra-linguistic formatting
  - Convert "non-standard" words like numbers, abbreviations, misspellings . . . into "normal" words
  - Misspelling correction
  - Abbreviation expansion (including novel abbreviations)
  - Possibly stemming
- May also involve some form of word sense disambiguation
  - Multiple identically spelled tokens with different properties, such as bass and bass: map them to bass

Where is normalization needed?

- Very little in cases like this:

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.
Where is normalization needed?

• A lot in cases like this:

```
CUST RCVD EFTTR CNRNG LOCAL SPVC
VISIT NECESSARY BUT CST STILL HAS P &C BELL SERV ON OLD TN AT RESIDENCE
ORDER CALLING CRDS PER CSR RQST
```

1st att. left msg for CBI from Lynda. will wait for call

547’s Sutton Plce Area Convertible 3BR 1400 SF 2BR, 2th, L-Shaped LR, S.E. Open Use, Gar, RF Ok, Mid $400K’s Thompson Kane Sta 319-8300


Humans are pretty good at this...
can you read this?

```
f u cn rd ths thn u r dng btr thn ny autmtc txt nrmztion prgrm cn do.
```

How about this?

Aocdrnig to a rscheearch at Cmabrigde Universtiy, it deosn’t mttaer in what order the ittters in a wrod are, the ylno tprmoetni ghnt is taht the trisf and tsaal rtteel be at the tghir eclap. The tser can be a lotat ssem and you can lltis daer it touthiw morbelp. Silt is ecuseab the nuamh dnim seed not daer yvee rtetel by fstiel, but the drow as a elohw.

Or this?

Goccdrnia to a hscheearcr at Emabrigdc Yinrvtisy, it teosn’dt rttaem in tawah rrdeo the stteerl in a drow are, the ylno tprmoetni ghnt is taht the trisf and tsaal rtteel be at the tghir eclap. The tser can be a lotat ssem and you can lltis daer it touthiw morbelp. Silt is ecuseab the nuamh dnim seed not daer yvee rtetel by fstiel, but the drow as a elohw.

Examples of Non-Standard Words (NSW)

• Numbers
  • 132 goats
  • 132 Park St
  • extension 132

• Acronyms/Letter sequences
  • NATO
  • UFO

• Abbreviations
  • Blvd.
  • St.
  • wbnpl (or wbf, wbfpl, wbfpl)

• Mixed examples
  • msdos, cdrom
  • named
  • w/11”, w/mahog.
Challenges with Text Normalization

• Genre/topic dependence
  • **named** is probably the ordinary word named in most cases;
    probably **name D** (= name daemon) in discussions of internet
domain servers
  • **BA** is probably bath(room) in real-estate classifieds;
    probably just **B A** in most other contexts
• Enumeration and selection
  • **BA**: bathroom, **B A**
  • **lv**: living (**Formal lv rm**), leave (**lv msg**)
    • (Think of this as a kind of pronunciation modeling problem)
• What to expand
  • Do you read **IMHO** as in **my humble opinion** or **I M H O**?

Agenda: Summary

• Discussion of HW5 (due date changed to Thursday)
• Questions, comments, concerns?
• Finish semantics discussion
• Text normalization
  • Tokenization
  • Abbreviations
  • Misspellings

... to be continued!