Words and their meaning

Three lectures:

- Collocations
  - multiple words together, different meaning than the sum of its parts
- Word disambiguation
  - one word, multiple meanings
- This time: Lexical Acquisition
  - verb subcategorization
  - attachment ambiguity
  - selectional preference
  - semantic similarity
    - multiple words, “same” meaning

Today’s Main Points

- What is Lexical Acquisition and why is it useful.
- Verb subcategorization.
- Attachment ambiguity
- Selectional preference
- Clustering words into semantically similar classes.

Lexical Acquisition

- Acquiring the properties of words
- Practical: filling holes in dictionaries
  - Lots of useful information isn’t in dictionaries anyway
    - e.g. “associated with” versus “associated to”
- Claim: most knowledge of language is encoded in words and their properties.
- Acquiring collocations and word sense disambiguation are examples of lexical acquisition, but there are many other types.

Why Lexical Acquisition

- Language evolves, i.e., new words and new uses of old words are constantly invented.
- Traditional Dictionaries were written for the needs of human users. Lexicons are dictionaries formatted for computers.
- In addition to the format, lexicons can be useful if they contain quantitative information. Lexical acquisition can provide such information.

Verb Phrase and Subcategorization

- Verb phrase consists of
  - Verb
    - a number of constituents
- Examples
  - VP → V disappear
  - VP → V NP prefer a morning flight
  - VP → V NP PP leave Boston in the morning
  - VP → V PP leave on Thursday
  - VP → V S said you had a $200 fare
  - Sentential complement
Different verbs, different constituents

- A verb phrase can have many possible kinds of constituents, but
- Not every verb is compatible with every verb phrase

**Examples**
- “want” VP → V NP: “I want a flight”
- “find” VP → V VPrep: “I want to fly to…”
- “find” VP → V VPrep: “I found a flight”

- **Transitive**, take a direct object
  - “find” VP → V NP
- **Intransitive**, do not take a direct object
  - “disappear” VP

**Transitive and Intransitive** are simple examples of verb subcategorization.

Verb Subcategorization

- Verbs express their semantic arguments with different syntactic means.
- “frame” = slots for arguments of the verb
- “category” = verbs that take the same semantic args e.g. verbs with semantic arguments theme and recipient
- “subcategory” = verbs that use the same syntactic means to express these semantic arguments.

**Additional examples:**
- “He donated a large sum of money to the church.”
- “He gave the church a large sum of money.”

Examples of subcategorization frames

- **Intransitive verb**
  - NP[subject]
  - “The woman walked.”
- **Transitive verb**
  - NP[subject] NP[object]
  - “John loves Mary.”
- **Ditransitive verb**
  - NP[subject], NP[direct object], NP[indirect object]
  - “Mary gave Peter flowers.”
- **Intransitive with PP**
  - NP[subject], PP
  - “I rent in Northampton.”
- **Sentential complement**
  - NP[subject], clause
  - “I know (that) she likes you.”
- **Transitive with sentential complement**
  - NP[subject], NP[object], clause
  - “She told me that Gary is coming.”

Subcategorization needed for parsing

- She told the man where Peter grew up.
- She found the place where Peter grew up.
- She told [the man] [where Peter grew up].
- She found [the place [where Peter grew up]].

Helps us get attachment right.

- Unfortunately most dictionaries don’t contain subcategorization frames, and those that do are horribly incomplete.

One verb, multiple subcategorizations

- One verb can take different subcategorization frames

**Example:** “find”
- VP → V NP ... find a flight
- VP → V NP NP ... find me a flight

Learning subcategorization frames

[Brent 1993]

- Does some particular verb take direct object frame VP → V NP?
- **Cues for frames**
  - e.g. assume that pattern “verb (pronoun | capitalized word) punctuation” identifies direct object frame with error rate e=0.1
- **Count occurrences**
  - n = number of occurrences of verb in question
  - m = number of occurrences of cue with verb
- **Hypothesis testing, H0 = verb does not take frame**
  $$ P(H0|\text{cue count} \geq m) = \sum_{r=m}^{n} \binom{n}{r} e^r (1-e)^{n-r} $$
Learning subcategorization frames
[Brent 1993] [Manning 1993]

- Brent's system does well at precision, but not well at recall.
- (Manning, 93)'s system addresses this problem by using a tagger and running the cue detection on the output of the tagger.
  - e.g. say “find/V DET NP” indicates direct object frame
- Manning’s method can learn a large number of subcategorization frames, even those that have only low-reliability cues.

### Learned subcategorization frames
[Manning 1993]

<table>
<thead>
<tr>
<th>Verb</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Oxford AL Dictionary</th>
</tr>
</thead>
<tbody>
<tr>
<td>bridge</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>burden</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>depict</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>emanate</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>leak</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>occupy</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>remark</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>retire</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Error in remark: attributed intransitive frame, probably due to “And here we are 10 years later with the same problems,” Mr. Smith remarked.

### Attachment Ambiguity

- Where to attach a phrase in the parse tree?
- “I saw the man with the telescope.”
  - What does “with a telescope” modify?
  - Is the problem AI complete? Yes, but…
  - Proposed simple structural factors
    - Right association [Kimball 1973]
    - ‘low’ or ‘near’ attachment = ‘early closure’ of NP
    - Minimal attachment [Frazier 1978]
      (depends on grammar) = ‘high’ or ‘distant’ attachment
      = ‘late closure’ (of NP)

Such simple structural factors dominated in early psycholinguistics, and are still widely invoked.
- In the V NP PP context, right attachment gets right 55-76% of the cases…
- But this means that it gets wrong 33-45% of the cases!

### Simple model

- (Log) likelihood ratio
  - A common and good way of comparing between two exclusive alternatives
  - Same idea as a naïve Bayes classifier

\[
\log \frac{P(\text{preposition} | \text{verb})}{P(\text{preposition} | \text{noun})}
\]

- if >0, attach to verb, if <0 attach to noun
- For example, P(with a spoon | ate) > P(with a spoon | cake)
Attachment, Problematic Example

- "Chrysler confirmed that it would end its troubled venture with Maserati."

<table>
<thead>
<tr>
<th>w</th>
<th>C(w)</th>
<th>C(w, with)</th>
</tr>
</thead>
<tbody>
<tr>
<td>end</td>
<td>5156</td>
<td>607</td>
</tr>
<tr>
<td>venture</td>
<td>1442</td>
<td>155</td>
</tr>
</tbody>
</table>

- Get wrong answer:
P(with|end) = (607/5156) = 0.118
P(with|venture) = (155/1442) = 0.107

- Should also express preference for attaching ‘low’.

Attachment Method [Hindle & Rooth 1993]

- Event space: all V NP PP* sequences but PP must modify V or first N
- Don’t directly decide whether PP modifies V or N
- Rather look at binary random variables
  - VA_p: is there a PP headed by p which attaches to v
  - NA_p: is there a PP headed by p which attaches to n
- Both can be 1:
  "He put the book on World War II on the table."

Attachment Method [Hindle & Rooth 1993]

- Independence assumptions
  \[ P(\text{VA}_p, \text{NA}_p | v, n) = P(\text{VA}_p | v, n) P(\text{NA}_p | v, n) \]
  \[ = P(\text{VA}_p | v) P(\text{NA}_p | n) \]
- Decision space: first PP after NP. [NB!]
  \[ P(\text{Attach}(p)=v|v,n) = P(\text{VA}_p=1, \text{NA}_p=0|v,n) \]
  \[ = P(\text{VA}_p=1|v) P(\text{NA}_p=0|n) \]
- It doesn’t matter what $\text{VA}_p$ is! If both are true, the first PP after the NP must modify the noun (in phrase structure trees, lines don’t cross).

Attachment Method [Hindle & Rooth 1993]

- But conversely, in order for the first PP headed by the preposition p to attach to the verb, both $\text{VA}_p=1$ and $\text{NA}_p=0$ must hold.
  \[ P(\text{Attach}(p)=v|v,n) = P(\text{VA}_p=1, \text{NA}_p=0|v,n) \]
  \[ = P(\text{VA}_p=1|v) P(\text{NA}_p=0|n) \]
- We assess which is more likely by a (log) likelihood ratio:
  \[ \lambda(v, n, p) = \log \frac{P(\text{Attach}(p)=v|v,n)}{P(\text{Attach}(p)=n|v,n)} \]
  \[ = \log \frac{P(\text{VA}_p=1|v, n)P(\text{NA}_p=0|v)}{P(\text{NA}_p=1|n)} \]
  \[ = \log \frac{\text{VA}_p=1|v, n)P(\text{NA}_p=0|v)}{P(\text{NA}_p=1|n)} \]
- If large positive, decide verb attachment; if large negative, decide noun attachment.

Attachment Method [Hindle & Rooth 1993]

- How do we learn probabilities?
  From (smoothed) MLEs:
  \[ P(\text{VA}_p=1|v) = \frac{C(v,p)}{C(v)} \]
  \[ P(\text{NA}_p=1|n) = \frac{C(n,p)}{C(n)} \]
- How do we get estimates from unlabeled corpus?
  Use partial parser, and look for unambiguous cases:
  - “The road to London is long and winding.”
  - “She sent him to the nursery to gather up his toys.”

Attachment Method [Hindle & Rooth 1993]

- Hindle and Rooth heuristically determine $C(v,p)$, $C(n,p)$ and $C(n,0)$ from unlabeled data:
  1. Build an initial model by counting all unambiguous cases.
  2. Apply initial model to all ambiguous cases and assign them to the appropriate count if I exceeds a threshold (2/-2).
  3. Divide the remaining ambiguous cases evenly between the counts (increase $C(v,p)$ and $C(n,p)$ by 0.5 for each).
Attachment Method Example
[Hindle & Rooth 1993]

- "Moscow sent more than 100,000 soldiers into Afghanistan..."

Other attachment issues

- There are attachment questions other than prepositional phrases
  - adverbial, participial, noun compounds
  - Examples
door bell manufacturer
[door bell] manufacturer
Unix system administrator
Unix [system administrator]
- Data sparseness is a bigger problem with many of these
- In general, indeterminacy is quite common
  - "We have not signed a settlement agreement with them."
  - Either reading seems equally plausible.

Lexical acquisition, semantic similarity

- Previous models give same estimate to all unseen events.
- Unrealistic - could hope to refine that based on semantic classes of words
- Examples
  - "Susan had never eaten a fresh durian before."
  - Although never seen "eating pineapple" should be more likely than "eating holograms" because pineapple is similar to apples, and we have seen "eating apples".

An application: selectional preferences

- Most verbs prefer arguments of a particular type. Such regularities are called selectional preferences or selectional restrictions.
- "Bill drove a..." Mustang, car, truck, jeep
- Selectional preference strength: how strongly does a verb constrain direct objects
- "see" versus "unknotted"

Measuring selectional preference strength

- Assume we are given a clustering of (direct object) nouns. Resnick (1993) uses WordNet.
  \[ S(v) = D(P(C|v)|P(C)) = \sum P(c|v) \log \frac{P(c|v)}{P(c)} \]
- Selectional association between a verb and a class
  \[ A(v, c) = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{S(v)} \]
  Proportion that its summand contributes to preference strength.
- For nouns in multiple classes, disambiguate as most likely sense:
  \[ A(v, n) = \max_{c \in \text{classes}(n)} A(v, c) \]

Selection preference strength
(made up data)

<table>
<thead>
<tr>
<th>Noun class c</th>
<th>P(eat)</th>
<th>P(see)</th>
<th>P(find)</th>
</tr>
</thead>
<tbody>
<tr>
<td>people</td>
<td>0.25</td>
<td>0.25</td>
<td>0.33</td>
</tr>
<tr>
<td>furniture</td>
<td>0.25</td>
<td>0.01</td>
<td>0.25</td>
</tr>
<tr>
<td>food</td>
<td>0.25</td>
<td>0.97</td>
<td>0.25</td>
</tr>
<tr>
<td>action</td>
<td>0.25</td>
<td>0.01</td>
<td>0.25</td>
</tr>
<tr>
<td>SPS S(v)</td>
<td>1.76</td>
<td>0.00</td>
<td>0.35</td>
</tr>
</tbody>
</table>

A(eat, food) = 1.08
A(find, action) = -0.13
Selectional Preference Strength example (Resnick, Brown corpus)

<table>
<thead>
<tr>
<th>Verb</th>
<th>Noun n</th>
<th>(\lambda(v,n))</th>
<th>Class</th>
<th>Noun n</th>
<th>(\lambda(v,n))</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>find</td>
<td>label</td>
<td>1.10</td>
<td>0.46</td>
<td>speech act</td>
<td>tragedy</td>
<td>0.46</td>
</tr>
<tr>
<td>hear</td>
<td>story</td>
<td>1.89</td>
<td>0.46</td>
<td>communication</td>
<td>communication</td>
<td>0.46</td>
</tr>
<tr>
<td>reply</td>
<td>event</td>
<td>1.31</td>
<td>0.46</td>
<td>statement</td>
<td>feature</td>
<td>0.46</td>
</tr>
<tr>
<td>read</td>
<td>article</td>
<td>1.23</td>
<td>0.46</td>
<td>communication</td>
<td>feature</td>
<td>0.46</td>
</tr>
<tr>
<td>see</td>
<td>friend</td>
<td>5.79</td>
<td>0.46</td>
<td>entity</td>
<td>method</td>
<td>0.01</td>
</tr>
<tr>
<td>write</td>
<td>letter</td>
<td>7.26</td>
<td>0.46</td>
<td>writing</td>
<td>market</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Andrew McCallum, UMass Amherst

But how might we measure word similarity for word classes?

- Vector spaces

A document-by-word matrix \(A\).

<table>
<thead>
<tr>
<th>cosmonaut</th>
<th>astronaut</th>
<th>moon</th>
<th>car</th>
<th>truck</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Similarity measures for binary vectors

- Matching coefficient: \(|X \cap Y| / |X| + |Y|\)
- Dice coefficient: \(2|X \cap Y| / (|X| + |Y|)\)
- Jaccard coefficient: \(|X \cap Y| / |X|\)
- Overlap coefficient: \(\min(|X|, |Y|) / \sqrt{|X||Y|}\)

Example of cosine measure on word-by-word matrix on NYT

Cosine measure

\[
\cos(X, Y) = \frac{X \cdot Y}{|X||Y|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2}}
\]

maps vectors onto unit circle by dividing through by lengths:

\[
|X| = \sqrt{\sum_{i=1}^{n} X_i^2}
\]

Focal word | Nearest neighbors |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gaffes</td>
<td>.732</td>
</tr>
<tr>
<td>list</td>
<td>.913</td>
</tr>
<tr>
<td>engineer</td>
<td>.758</td>
</tr>
<tr>
<td>simple</td>
<td>.964</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sauce</td>
<td>.726</td>
<td>.726</td>
</tr>
<tr>
<td>pepper</td>
<td>.726</td>
<td>.726</td>
</tr>
<tr>
<td>salt</td>
<td>.726</td>
<td>.726</td>
</tr>
<tr>
<td>stop</td>
<td>.726</td>
<td>.726</td>
</tr>
<tr>
<td>drop</td>
<td>.726</td>
<td>.726</td>
</tr>
<tr>
<td>drop</td>
<td>.579</td>
<td>.579</td>
</tr>
<tr>
<td>engineer</td>
<td>.591</td>
<td>.591</td>
</tr>
<tr>
<td>research</td>
<td>.598</td>
<td>.598</td>
</tr>
<tr>
<td>drug</td>
<td>.685</td>
<td>.685</td>
</tr>
<tr>
<td>always</td>
<td>.808</td>
<td>.808</td>
</tr>
<tr>
<td>things</td>
<td>.963</td>
<td>.963</td>
</tr>
<tr>
<td>you</td>
<td>.963</td>
<td>.962</td>
</tr>
</tbody>
</table>
Probabilistic measures

(Dis-)similarity measure | Definition
--- | ---
KL divergence | $D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$
Skew | $D(q\|\alpha r + (1 - \alpha)q)$
Jensen-Shannon (was IRad) | $\frac{1}{2}D(p\|\frac{1 + \alpha}{2} r) + D(q\|\frac{1 + \alpha}{2} p)$
$L_1$ norm (Manhattan) | $\sum_i |p_i - q_i|$