In today's lecture
Given any sentence e.g. “British Left waffles on Falklands”
Many different parses
Many different semantic readings
Could we make a shallow first guess to save time?
Can we tag every word with some initial information to aid higher level processes?
... and while we're at it -
What do we mean when we say a word has meaning?

The problem
As we discussed last week many NLP processes are serial in nature:
• Sentence interpretation
  “The boy saw the girl with the telescope”
  Morphology -> Syntax -> Semantics -> Pragmatics
• Every aspect of language is ambiguous
• What happens when we make the wrong decision early on?
  Breadth first versus depth first search
  (space versus time)
Could accurate guessing early on avoid work?

Tagging
There are several things we can guess without any further analysis
POS tags
Named entities in the text
Noun Phrases
Verb arguments
All of these can be guessed quickly with a reasonable level of accuracy
Many applications will not require any further NLP processing
Index generation
Information Retrieval
Automatic Summarization & Glossing
What about guessing word meaning?

What do words mean?
• Compositional Nat. Lang. semantics
  words contribute content to a sentence
  no “meaning” outside of a sentence (cannot be true or false)
  Lexicon is a listing of symbolic fragments without internal structure
• Word definitions are more complex than this!
  Word definitions have internal structure
  Word definitions link to other definitions
• Words have different meanings (senses)
  Homonyms - same word - different (unrelated meanings)
  Homographs/phones - same form/same sound
• Meanings can be expressed in different words
  Synonyms - different words - same meaning (sense)
  Strong versus Weak synonyms - “Big” versus “Large”
  Different registers - “Strong” versus “Powerful”

Words and Dictionaries
• What are dictionaries for?
  right adj. located nearer the right hand. esp. being on the right when facing the same direction as the observer.
  left adj. located nearer to this side of the body than the right.
  red n. the colour of blood or a ruby.
  blood n. the red liquid that circulates in the heart, arteries and veins of animals.
Claims
• Dictionaries do not define words.
  Rather they disambiguate words
• Words are related to each other.
How are dictionaries written?

- Intuition
- Corpora & Concordances

What does “make” mean? (have a look in the corpus)

However, this did not make the website a bad one. It enabled me to make an opinion about which various other colours to make headings etc. stand out. Text that I wouldn’t make it one big section (see http://www.webcorp.org.uk/)

Word senses

John went to the bank.

Bank as noun

1. financial institution
2. sloping land (esp. beside water)
3. supply or stock held in reserve (blood bank)
4. Bank (a building for banking)
5. an arrangement of similar objects in a row
6. money box a container for coins (piggy bank)
7. a long ridge
8. funds held by a gambling house
9. cart, camber a slope in the turn of a road
10. a flight maneuver

(from Wordnet 3.0)

Polysemy

- Homographs
  - Unrelated senses
    - (river)bank/(money)bank
- Polysemy
  - Related but distinct senses
    - John baked a cake/John baked a potato
    - the ceiling was high/the bookcase was high
- Vagueness versus Polysemy
  - literal sense -> metaphor -> polysemous new sense
- Tests of polysemy
  - Etymology (but consider “pupil”)
  - Zeugma “John expired the same day as his tv licence”

Zipf’s law(s) Principle of least effort

- Zipf’s law
  - If words are ranked according to their frequency then
    - Frequency inversely proportional to Rank
    \[ F \sim \frac{1}{R} \]
  - Implication: corpora are never big enough to cover all words.
- Zipf’s law applied to word senses
  - Relationship of frequency to number of word senses is
    \[ \text{Senses} \sim \frac{1}{\text{Rank}} \]
  - Word Rank: 10,000/about 2.1 senses; 5,000 rank 3 senses; 2,000 rank 4.6 senses
  - Implications:
    - Common words are very ambiguous
    - Rare words are not but little training data available

Lexicographer intuitions

Commercial reasons

More data - more intuitions!

But:

Concordance data supports Zipf’s law of word senses (roughly)

Frequency allows for a greater degree of sense extension.

More opportunity for a word to be used in a new sense

Wordnet definition of verb “make” has 48 senses.

Reasons for ambiguity

Lexicographer intuitions

Commercial reasons

More data - more intuitions!

But:

Concordance data supports Zipf’s law of word senses (roughly)

Frequency allows for a greater degree of sense extension.

More opportunity for a word to be used in a new sense

Wordnet definition of verb “make” has 48 senses.

Machine Translation

- Words are not consistently cross-language ambiguous.

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>one word/several senses</td>
<td>several words/single senses</td>
</tr>
<tr>
<td>drug (american english)</td>
<td>drogue, medicant</td>
</tr>
<tr>
<td>room</td>
<td>salle chambre piece</td>
</tr>
<tr>
<td>several word/single senses</td>
<td>one word/several senses</td>
</tr>
<tr>
<td>railing/hurdle/gate/picket</td>
<td>grille</td>
</tr>
</tbody>
</table>

- Decisions have to be made in machine translation

“Little John was looking for this toy box. Finally he found it. The box was in the pen. John was very happy” (Bar-hillel, 1960)
Other Applications of WSD

• Speech generation
  ambiguous words can be pronounced differently.
  lead (pipes)
  lead (dog)

• Speech understanding
  can be used to detect homophones
  hoof/hoot & ceiling/sealing

• Information Retrieval
  manual disambiguated a IR test corpus and found improved performance by 2%
  Complex Queries tend to self-disambiguate (Kowetz & Croft, 1992)
  High quality WSD - small increase in performance
  Poor quality WSD - large decrease in performance
  (Sandstrom, 1994)

Selectional Restrictions

• Disambiguation based on selectional restrictions.
• Each sense in the lexicon has a formula of features
• 100 semantic features (HUMAN/ANIMATE/ABSTRACT...)
• Formulae relate features with word & neighboring words
  Nouns: appropriate semantic feature
  Verbs: preferred semantic features of noun(s) it governs
  Adverbs: preferred semantic feature of verb
• Polysemy handled by associating several related formula to a single sense

Disambiguation using preference semantics

“The adder drank from the pool”
Drink [+subject(animate) +object(liquid) +object(edible)]
Adder - calculator (inanimate)
Adder - snake (animate)
• If no semantic conflicts preferences are loosened
“My car drinks gasoline.”
car (+machine)
gasoline (+liquid, -edible)
• New formula added as a polysemous sense of drink
• Use of pseudo texts (frames) to understand usage.

Problems

• Often semantic restrictions are broken/ignored in text
  “But it fell apart in 1931, perhaps because people realized you can’t eat gold for lunch if you’re hungry” (WSJ corpus)
• Different degrees of restrictiveness
  Drink versus Make
• Requires syntactic analysis of text before WSD
  Hard to do
  May not be required for the rest of the application
• Manual construction of formulae
  Knowledge acquisition bottleneck

Dictionary based approaches

• 1980s Machine readable dictionaries became available.
• Offer a way out of the knowledge acquisition bottleneck
  But
• More senses offered than previous systems had assumed.
• WSD - disambiguate relative to MRD senses.
  But do dictionaries offer the right kinds of information to aid task?
  (rather than just enumerate senses)
• Typically Learner dictionaries are used
  Cobuild/OALD/LDOCE etc.

Lesk (1986)

Key Idea
the sense of a word is linked to senses of the words surrounding the word in the text
• use textual definitions of words from OALD
• classify each sense by the words used in its definition

Given a text
• disambiguate each word by looking at a ten word context window
• choose the sense which shares the most words in its definition with words in the context window.

MRD & WSD

Lesk (1986)

Key Idea
the sense of a word is linked to senses of the words surrounding the word in the text
• use textual definitions of words from OALD
• classify each sense by the words used in its definition

Given a text
• disambiguate each word by looking at a ten word context window
• choose the sense which shares the most words in its definition with words in the context window.
John found a *pine cone*.

Pine
1. waste away from sorrow or illness
2. kind of evergreen tree with needle-shaped leaves

Cone
1. solid body which narrows to a point
2. something of this shape whether solid or hollow
3. fruit of certain evergreen trees

**Notes**
Lesk found shrinking the window had little effect on accuracy. 50-70% disambiguate short samples of Pride and Prejudice (est.).

---

**Simulated Annealing**

Improve technique by considering all words in a sentence
- Very computationally expensive to do

Simulated Annealing: "Heat metal then strengthen by cooling."

1. Initial guess at solution
2. generate permutations & evaluate which are improvements
3. To avoid local minima: stochastic random element in generation depending on temperature of solution.
   - High temperature - more random
4. Gradually cool solution till settles on final solution.
   - Not guaranteed to find optimal solution but ...
   - Computationally efficient

**Algorithm**

- Given a sentence with N words, represent senses of W as
  \[ S_1, S_2, \ldots, S_N \]
- Initial Configuration C
  - Choose \( S_i \) for all i. Combine all words in all sense definitions in List L
  - If R be the sum of redundancy (repetition) of words in List L
    - Energy \( E = \frac{1}{1+R} \)
  - Randomly choose a new sense in one word in C and create C'
    - Change to C' if \( E' < E \) OR
      - on probability \( P \) where temperature T is a constant and
        \[ P = e^{-\frac{\Delta E}{T}} \]
- Repeat 1000 times
  - Finish if no change in configuration if not repeat with T = .9 T

---

**Evaluation**

- Used LDOCE word senses
- Senses are frequency ordered
- Senses are homographically sorted

from 50 examples
- 47% correct assigned LDOCE senses
- 72% correct assigned homograph senses

**Advantages**
- No syntax/POS guidance
- Only used word counts from definitions

---

**Statistical Approaches**

- Statistical NLP techniques are currently very trendy...
  - greater availability of machine readable corpora
  - big increase in computer power
  - better understanding of mathematical foundations

But not a new idea...
- Masterman (1957) MT using Roget's thesaurus
  (but not implemented on large scale)
- Spark-Jones (1964) automatically creating thesaurus-like structures from corpora (not implemented - only suggested)

---

**Two Claims about Senses**

**Claim 1:**
One Sense Per Discourse (Gale, 1992)
Strong tendency (98%) that identical words will have the same sense in the discourse.
If so, pull out all instances of word in text and do global WSD
More evidence to make decision
Selectional restrictions (at sentence level) will be less useful.
Knowles (1998) only true if dealing homograph senses

**Claim 2:**
One Sense Per Collocation (Yarowsky, 1993)
Different senses are unlikely to share the same collocation of surrounding words.
This is the basic assumption of all statistical WSD techniques.
Tagging with Thesaurus Categories

- Yarowsky (1992) sense tagging using Roget

**Key Idea**
Build statistical models of each category in Roget from raw untagged text
disambiguate new words using these statistical models

- Large corpus Grolliers Encyclopedia 10 million words
- 100 word window of context
- Roget's categories
  15 classes (Science & Technology/The body and senses...)
  1073 sub-classes

Building a Statistical Model

**Training**
For every sub-category (SC)
For every word w in (SC)
  Extract all 100 word windows of context from corpus
  For each word w in all windows for SC calculate
    \[
    P(w_i | SC) = \frac{1}{\text{count}(w_i, SC)}
    \]

**Tagging**
Given an ambiguous word
- Examine 100 word context of ambiguous word
- Assume \( Pr(SC) \) is equal for all categories
- \( Pr(w) \) is constant for entire text

**Results**
- 12 words - 97% accuracy
- good on nouns - okay on verbs and poor on adjectives and adverbs

LDOCE

Longman’s Dictionary of Contemporary English (1978)

Learner’s Dictionary
Core Vocab. of 2000 words
No attempt to control examples of usage

Three-level structure of word senses
Homograph level
Fine-grained (polysemy) level sub-sense

Each sense is given a homograph and sense number - 1_2a etc.
syntax information [T1] etc.

WordNet

Miller (1986) Cognitive Model of Human mental lexicon

Basic Block is the SynSet (Synonym Set)
groups of words with "same" meaning

Organized as a semantic net:
  antonym/hyponym/meronym links
  55% of synsets also have a gloss
  Size
    Nouns 94474 forms | 16317 senses
    Verbs 10319 forms | 22066 senses
    Adjectives 20170 forms | 29881 senses
    Adverbs 4546 forms | 5677 senses
  Euro-Wordnet, Korean-Wordnet etc.
  SemCor - the Brown Corpus handtagged for WN senses
An Example

The Bank Model (Kilgariff (1997))

“John went to the bank to pay a bill”

Is this a typical case?

- People don’t agree when hand-tagging
- People don’t agree on how many senses a word has
- Dictionaries agree on how many senses a word has
- Multiple activation of senses
- Is sense enumeration realistic?

Alternative

Minimal definition plus rules of inference

Several proposals but this hasn’t been successfully implemented

Further Issues

- Limitations of statistical approaches
  - Corpora
  - Applications of unsupervised WSD

- Are dictionaries the right kind of knowledge source?
  - Information for disambiguation?
    - Granularity
    - Dictionary agreement on sense inventories
    - Human agreement on WSD?

- Bilingual dictionaries
  - Bilingual corpora (is this proper sense tagging?)

Further Reading

- A. Kilgariff (1999) “I don’t believe in Word Senses” IJCAI 97-12 (see pointer from module webpage) University of Brighton (1999) (part of the reading list for one of my suggested essays)