Word Sense Disambiguation

FSNLP, chapter 7

Christopher Manning and
Hinrich Schütze

Word sense disambiguation

- The task is to determine which of various senses of a word are invoked in context:
  - *the seed companies cut off the tassels of each **plant**, making it male sterile*
  - *Nissan's Tennessee manufacturing **plant** beat back a United Auto Workers organizing effort with aggressive tactics*
- This is an important problem: Most words are ambiguous (have multiple senses)
- Converse: words or senses that mean (almost) the same: *image, likeness, portrait, facsimile, picture*

WSD: Many other cases are harder

- *title:*
  - Name/heading of a book, statute, work of art or music, etc.
  - Material at the start of a film
  - The right of legal ownership (of land)
  - The document that is evidence of this right
  - An appellation of respect attached to a person's name
  - A written work

WSD: The many meanings of *interest* [n.]

- Readiness to give attention to or to learn about something
- Quality of causing attention to be given
- Activity, subject, etc., which one gives time and attention to
- The advantage, advancement or favor of an individual or group
- A stake or share (in a company, business, etc.)
- Money paid regularly for the use of money

WSD: types of problems

- Homonymy: menaings are unrelated: *bank* of river and *bank* financial institution
- Polysemy: related meanings (as on previous slides)
- Systematic polysemy: standard methods of extending a meaning, such as from an organization to the building where it is housed.
- A word frequently takes on further related meanings through systematic polysemy or metaphor

Word sense disambiguation

- Most early work used semantic networks, frames, logical reasoning, or "expert system" methods for disambiguation based on contexts (e.g., Small 1980, Hirst 1988).
- The problem got quite out of hand:
  - The word expert for 'throw' is "currently six pages long, but shouw be ten times that size" (Small and Rieger 1982)
- Supervised sense disambiguation through use of context is frequently extremely successful – and is a straightforward classification problem
- "You shall know a word by the company it keeps" – Firth
- However, it requires extensive annotated training data

Some issues in WSD

- Supervised vs. unsupervised
  - Or better: What are the knowledge sources used?
- Pseudowords
  - Pain-free creation of training data
  - Not as realistic as real words
- Upper and lower bounds: how hard is the task?
  - Lower bound: go with most common sense (can vary from 20% to 90+% performance)
  - Upper bound: usually taken as human performance

A simple but OK approach: Naive Bayes WSD

- $\vec{x}$ is our context (something like a 100 word window)
- c_k is a sense of the word

$$\begin{aligned}
\text{Choose } c' &= \arg\max_{c_k} P(c_k|\vec{x}) \\
&= \arg\max_{c_k} \frac{P(\vec{x}|c_k)}{P(\vec{x})} P(c_k) \\
&= \arg\max_{c_k} P(\vec{x}|c_k) P(c_k) \\
&= \arg\max_{c_k} [\log P(\vec{x}|c_k) + \log P(c_k)] \\
&= \arg\max_{c_k} \left[\sum_{v_j \text{ in } \vec{x}} \log P(v_j|c_k) + \log P(c_k) \right]
\end{aligned}$$

Other (semi-)supervised WSD

- Brown et al. (1991): used just one key indicating (linguistic) feature (e.g., object of verb) and partitioned its values
- Lesk (1986) used a dictionary; Yarowsky (1992) used a thesaurus
- Use of a parallel corpus (Brown et al. 1991) or a bilingual dictionary (Dagan and Itai 1994)
- Use of decomposable models (a more complex Markov random field model) (Bruce and Wiebe 1994, 1999)

Lesk (1986)

- Words in context can be mutually disambiguated by overlap of their defining words in a dictionary
  - **ash**
    1. the **solid** residue left when **combustible** material is thoroughly **burned** ...
    2. Something that symbolizes grief or repentence
  - **coal**
    1. a black or brownish black **solid combustible** substances ...
- We'd go with the first sense of **ash**
- Lesk reports performance of 50%–70% from brief experimentation

Collocations/selectional restrictions

- Sometimes a single feature can give you very good evidence – and avoids need for feature combination
- Traditional version: selectional restrictions
  - Focus on constraints of main syntactic dependencies
  - *I hate washing **dishes***
  - *I always enjoy spicy **dishes***
  - Selectional restrictions may be weak, made irrelevant by negation or stretched in metaphors or by odd events
- More recent versions: Brown et al. (1991), Resnik (1993)
  - Non-standard good indicators: tense, adjacent words for collocations (*mace spray*; *mace* and *parliament*)

Global constraints: Yarowsky (1995)

- One sense per discourse: the sense of a word is highly consistent within a document
  - True for topic dependent words
  - Not so true for other items like adjectives and verbs, e.g. *make, take*
  - Krovetz (1998) "More than One Sense Per Discourse" argues it isn't true at all once you move to fine-grained senses
- One sense per collocation: A word reoccurring in collocation with the same word will almost surely have the same sense
  - This is why Brown et al.'s (1991) use of just one disambiguating feature was quite effective

Unsupervised disambiguation

- Word sense discrimination (Schütze 1998) or clustering
- Useful in applied areas where words are usually used in very specific senses (commonly not ones in dictionaries!). E.g., *water table* as bit of wood at bottom of door
- One can use clustering techniques
- Or assume hidden classes and attempt to find them using the EM (Expectation Maximization) algorithm – covered later – (Schütze 1998)

WSD: Senseval competitions

- Senseval 1: September 1998. Results in *Computers and the Humanities* 34(1–2). OUP Hector corpus.
- Senseval 2: in first half of 2001. WordNet senses.
- Sense-tagged corpora available:
  - http://www.itri.brighton.ac.uk/events/senseval/
- Comparison of various systems, all the usual suspects (naive Bayes, decision lists, decomposable models, memory-based methods), and of foundational issues

WSD Performance

- Varies widely depending on how difficult the disambiguation task is
- Accuracies of over 90% are commonly reported on some of the classic, often fairly easy, word disambiguation tasks (*pike*, *star*, *interest*, ...)
- Senseval brought careful evaluation of difficult WSD (many senses, different POS)
- Senseval 1: more fine grained senses, wider range of types:
  - Overall: about 75% accuracy
  - Nouns: about 80% accuracy
  - Verbs: about 70% accuracy

What is a word sense?

- Particular ranges of word senses have to be distinguished in many practical tasks, e.g.:
  - translation
  - IR
- But there generally isn't *one* way to divide the uses of a word into a set of non-overlapping categories. Dictionaries provide neither consisentency nor non-overlapping categories usually.
- Senses depend on the task (Kilgarriff 1997)

Similar 'disambiguation' problems

- Sentence boundary detection
- *I live on Palm Dr. Smith lives downtown.*
- Only really ambiguous when:
  - word before the period is an abbreviation (which can end a sentence – not something like a title)
  - word after the period is capitalized (and can be a proper name – otherwise it must be a sentence end)
- Can be treated as 'disambiguating' periods (as abbreviation mark, end of sentence, or both simultaneously [haplology])

Similar 'disambiguation' problems

- Context-sensitive spelling correction:
- *I know their is a problem with there account.*