Word Sense Disambiguation

FSNLP, chapter 7

Christopher Manning and Hinrich Schütze

© 1999–2004

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: Many other cases are harder

- *modest:*
  - In evident apprehension that such a prospect might frighten off the young or composers of more modest_1 forms –
  - Tort reform statutes in thirty-nine states have effected modest_9 changes of substantive and remedial law
  - The modest_9 premises are announced with a modest and simple name –
  - In the year before the Nobel Foundation belatedly honoured this modest_0 and unassuming individual,
  - LinkWay is IBM's response to HyperCard, and in Glasgow (its UK launch) it impressed many by providing colour, by its modest_9 memory requirements,
  - In a modest_1 mews opposite TV-AM there is a rumpled hyperactive figure
  - He is also modest_0: the "help to" is a nice touch.

WSD: types of problems

- Homonymy: meanings 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

Unsupervised and semi-supervised WSD

- Really, if you want to be able to do WSD in the large, you need to be able to disambiguate all words as you go.
- You're unlikely to have a ton of hand-built word sense training data for all words.
- Or you might: the OpenMind Word Expert project:
  - http://teach-computers.org/word-expert.html

Unsupervised and semi-supervised WSD

- Main hope is getting indirect supervision from existing broad coverage resources:
  - Lesk (1986) used a dictionary; Yarowsky (1992) used a thesaurus
  - Use of a parallel corpus (Brown et al. 1991b) or a bilingual dictionary (Dagan and Itai 1994)

  This can be moderately successful. (Still not nearly as good as supervised systems. Interesting research topic.
- There is work on fully unsupervised WSD
  - This is effectively word sense clustering or word sense discrimination (Schütze 1998).
  - Usually no outside source of truth
  - Can be useful for IR, etc. though

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 (1991b) 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 (Schütze 1998)

WSD: Senseval competitions

- Senseval 1: September 1998. Results in *Computers and the Humanities* 34(1–2). OUP Hector corpus.
- Senseval 2: first half of 2001. WordNet senses.
- Senseval 3: first half of 2004. 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.*

Text categorization

- Have some predefined categories for texts
  - Predefined categories for news items on newswires – Reuters categories
  - Yahoo! classes (extra complexity: hierarchical)
  - Spam vs. not spam
- Word sense disambiguation can actually be thought of as text (here, context) categorization
  - But many more opportunities to use detailed (semi-) linguistic features

(Multinomial) Naive Bayes classifiers for WSD

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

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

- An effective method in practice, but also an example of a structure-blind 'bag of words' model

Senseval 2 results

- The hacked Naive Bayes classifier has no particular theoretical justification. One really cannot make sense of it in terms of the independence assumptions, etc., usually invoked for a Naive Bayes model
- But it is linguistically roughly right: nearby context is often very important for WSD: noun collocations (*complete **accident***), verbs (***derive** satisfaction*)
- In Senseval 2, it scores an average accuracy of 61.2%
- This model was just a component of a system we entered, but alone it would have come in 6th place out of 27 systems (on English lexical sample data), beating out all the systems on the previous slide

Disambiguating using 'language' models

- Supervised training from hand-labeled examples
- Train *n*-gram language model for examples of each sense, treating examples as a 'language'
  - estimate $P(\text{frog}|\text{large}, \text{green})$, etc.
  - reduce parameters by backing off where there is insufficient data: use $P(\text{frog}|\text{green})$ or $P(\text{frog})$
- Disambiguate based on in which 'language' the sentence would have highest probability
- Multinomial Naive Bayes models are class-conditional unigram language models
- Higher oder models can give some of the advantages of wide context bag of words models (Naive Bayes-like) and use of local structural cues around word

WSD methods

- One method: A multinomial naive Bayes classifier, add $\frac{1}{10}$ smoothing. Except words near the ambiguous word are weighted by a strongly peaked function (distance 3–5, $3\times$; distance 2, $5\times$, distance 1, $15\times$)
- Other methods (Senseval 2 entries):
  - Bagged decision trees with unigram, bigram, and long distance bigram features
  - Weighted vote of DT, NB, and *k*NN classifiers over short and long distance bigram features
  - Hierarchical LazyBoosting over large and small window bag-of-word features, and WordNet features
  - Support vector machine with IDF feature weighting

Naive Bayes models

- The *Naive Bayes assumption* is that the attributes used for description are all conditionally independent:

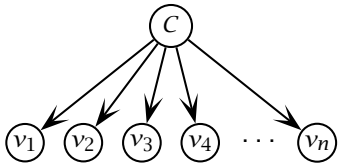
  Naive Bayes assumption

  $P(\vec{x}|c_k) = P(\{v_j|v_j \text{ in } \vec{x}\}|c_k) = \prod_{v_j \text{ in } \vec{x}} P(v_j|c_k)$
- This is commonly referred to as the *bag of words* assumption
- **Decision rule for Naive Bayes**

  Decide c' if $c' = \arg\max_{c_k} [\log P(c_k) + \sum_{v_j \text{ in } \vec{x}} \log P(v_j|c_k)]$
- Note that there are two Naive Bayes models (McCallum and Nigam 1998)

Two Naive Bayes models: Multinomial

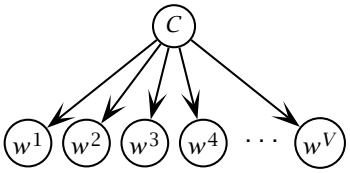
- v_j is word j of the context



- Model of Gale et al. (1992) (for WSD). Usual in StatNLP.
- The CPT for each multinomial is identical (tied parameters)
- The multinomial is estimated over the whole vocabulary.

Two Naive Bayes models: Bernoulli

- w^j is word (type) j of the vocabulary of features



- Each feature is binary yes/no (though could be count/range)
- Model normally presented in the graphical models literature
- Generally (but not always) performs worse
- Requires careful and aggressive feature selection

Naive Bayes models

- Feature selection: commonly count, χ^2 or mutual information, but there are methods to find non-overlapping features (Koller and Sahami 1996). Only important/relevant in Bernoulli model.
- Naive Bayes is simple, but often about as good as there is (Friedman 1997; Domingos and Pazzani 1997)
- There are successful more complex probabilistic classifiers, particularly TAN – Tree Augmented Naive Bayes (van Rijsbergen 1979; Friedman and Goldszmidt 1996)
- One can get value from varying context size according to type of word being disambiguated (commonly: noun is big context, verb is small context)

'Typical' McCallum and Nigam (1998) result: Reuters Money-FX category