Creating annotated corpora for supervised sense disambiguation

Lecture Two

Christiane Fellbaum

The lexical bottleneck

Supervised approaches require human-annotated “gold standard” text corpora (semantic concordances)

Text tokens of open-class words (nouns, verbs, adjectives, adverbs) are linked to specific entries (context-appropriate senses) in a lexical resource

Very expensive to produce, time consuming

Annotators don’t always agree with one another

Semantic annotation

Determine the meaning of (polysemous) words in their contexts

He asked the waiter for the ***check*** (bill)

She cashed the ***check*** (bank check)

I'll ***check*** *the door lock* (control, verify)

Select context-appropriate sense in a reference dictionary

Task requires word sense disambiguation for each text token

Why did she buy the paper?

A representative entry (noun "paper") in WordNet:

- # (n) paper (a material made of cellulose pulp derived mainly from wood or rags or certain grasses)
- # S: (n) composition, paper, report, theme (an essay (especially one written as an assignment)) "he got an A on his composition"
- # S: (n) newspaper, paper (a daily or weekly publication on folded sheets; contains news and articles and advertisements) "he read his newspaper at breakfast"
- # S: (n) paper (a medium for written communication) "the notion of an office running without paper is absurd"
- # S: (n) paper (a scholarly article describing the results of observationsor stating hypotheses) "he has written many scientific papers"
- # S: (n) newspaper, paper, newspaper publisher (a business firm that publishes newspapers) "Murdoch owns many newspapers"
- # S: (n) newspaper, paper (the physical object that is the product of a newspaper publisher) "when it began to rain he covered his head with a newspaper"

Challenge

Recall: the most frequently used word forms are also the most polysemous

What is annotated

Open-class words

(Why don't we annotate closed class words? Should we? Which ones?)

Multi-word units with atomic meaning

--phrasal verbs (*check out, check up*)

--opaque compounds (*road rage*)

--idioms (*hit the ceiling*)

WordNet

Most commonly used digital lexical resource

(Fellbaum 1998)

(more on WordNet to come later)

Cut-out of WordNet’s entry for *check*

13 noun senses

5 verb senses

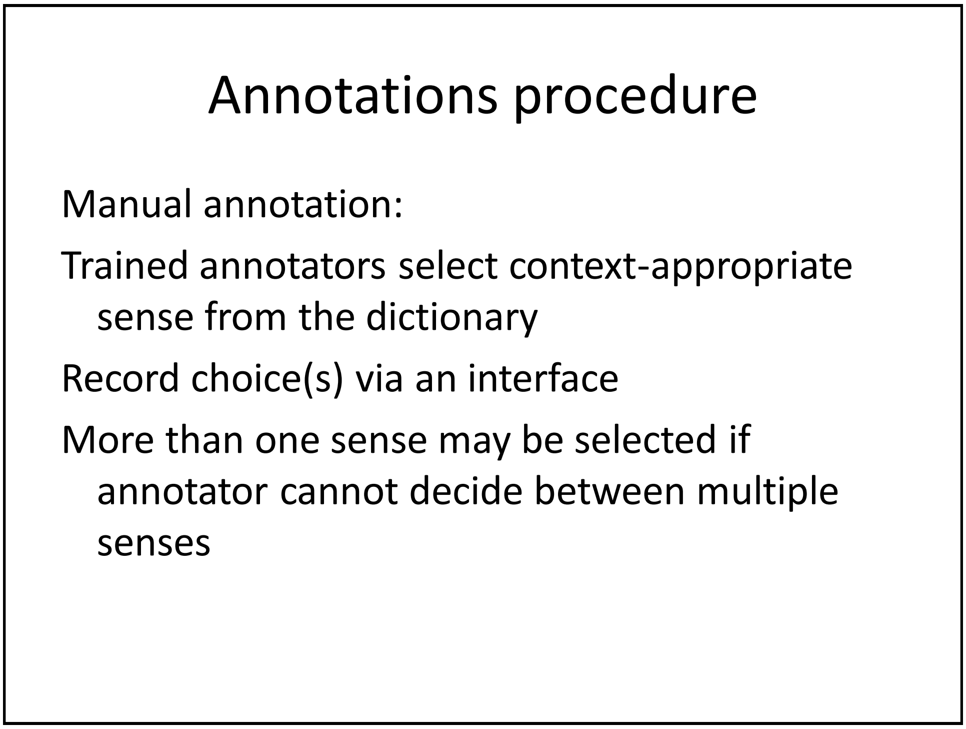
Noun

* S: (n) check, bank check, cheque (a written order directing a bank to pay money) "he paid all his bills by check"
* S: (n) assay, check (an appraisal of the state of affairs) "they made an assay of the contents"; "a check on its dependability under stress"
* S: (n) check, chit, tab (the bill in a restaurant) "he asked the waiter for the check"
* S: (n) arrest, check, halt, hitch, stay, stop, stoppage (the state of inactivity following an interruption) "the negotiations were in arrest"; "hel
* S: (n) confirmation, verification, check, substantiation (additional proof that something that was believed (some fact or hypothesis or theor
* S: (n) check, checkout, check-out procedure (the act of inspecting or verifying) "they made a check of their equipment"; "the pilot ran thro
* S: (n) check mark, check, tick (a mark indicating that something has been noted or completed etc.) "as he called the role he put a check ma
* S: (n) hindrance, hinderance, deterrent, impediment, balk, baulk, check, handicap (something immaterial that interferes with or delays acti
* S: (n) check, chip (a mark left after a small piece has been chopped or broken off of something)
* S: (n) check (a textile pattern of squares or crossed lines (resembling a checkerboard)) "she wore a skirt with checks"
* S: (n) bridle, check, curb (the act of restraining power or action or limiting excess) "his common sense is a bridle to his quick temper"
* S: (n) check (obstructing an opponent in ice hockey)
* S: (n) check ((chess) a direct attack on an opponent's king)

Verb

* S: (v) check, check up on, look into, check out, suss out, check over, go over, check into (examine so as to determine accuracy, quality, or co
* S: (v) check (make an examination or investigation) "check into the rumor"; "check the time of the class"
* S: (v) see, check, insure, see to it, ensure, control, ascertain, assure (be careful or certain to do something; make certain of something) "He
* S: (v) control, hold in, hold, contain, check, curb, moderate (lessen the intensity of; temper; hold in restraint; hold or keep within limits) "m
* S: (v) check (stop for a moment, as if out of uncertainty or caution) "She checked for an instant and missed a step”

Etc.



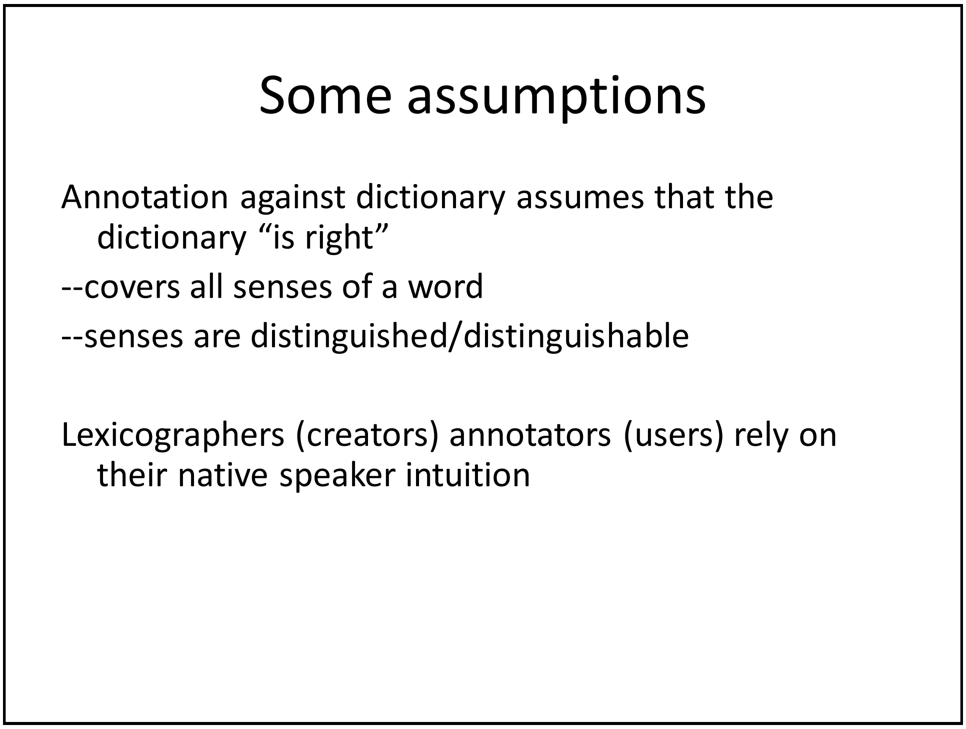
Annotations procedure

Manual annotation:

Trained annotators select context-appropriate sense from the dictionary

Record choice(s) via an interface

More than one sense may be selected if annotator cannot decide between multiple senses



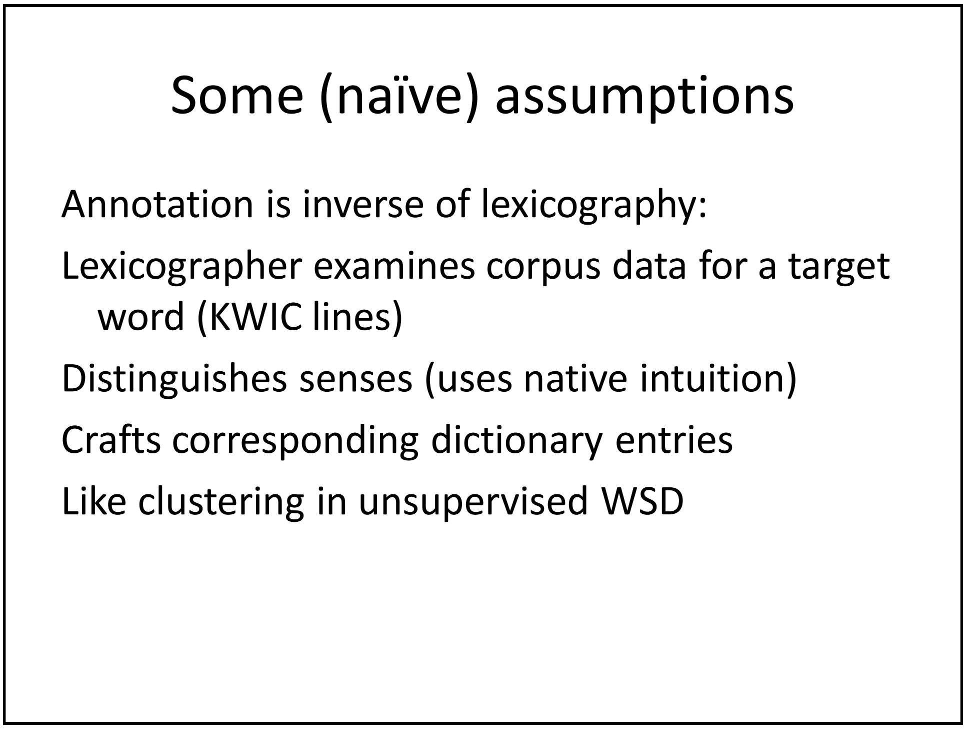
Some assumptions

Annotation against dictionary assumes that the dictionary “is right”

--covers all senses of a word

--senses are distinguished/distinguishable

Lexicographers (creators) annotators (users) rely on their native speaker intuition



Some (naïve) assumptions

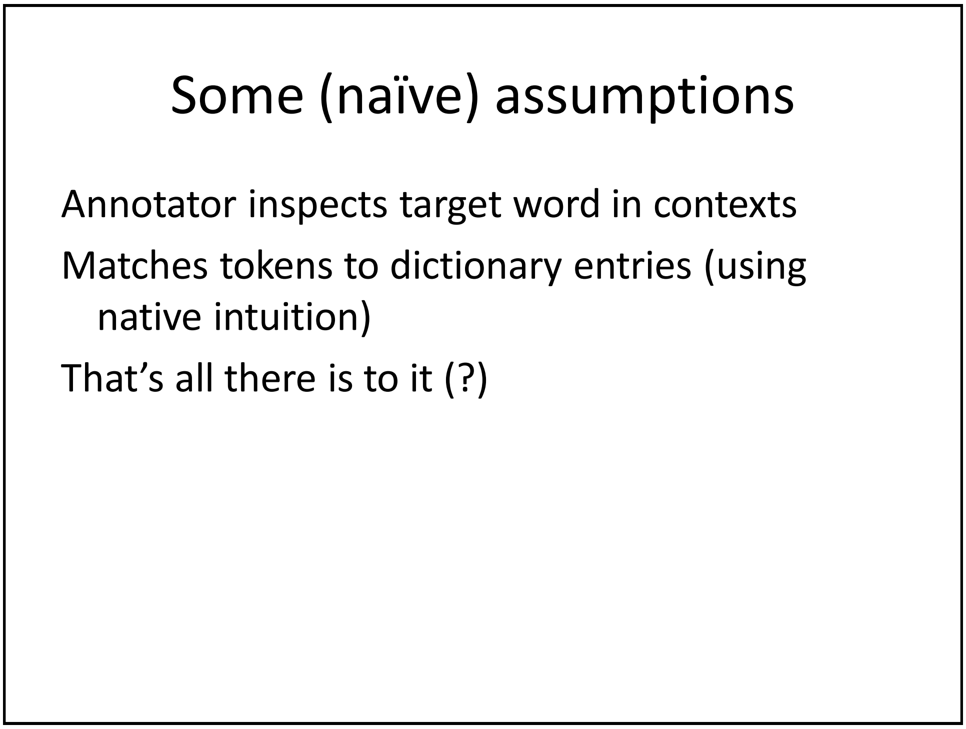
Annotation is inverse of lexicography:

Lexicographer examines corpus data for a target word (KWIC lines)

Distinguishes senses (uses native intuition)

Crafts corresponding dictionary entries

Like clustering in unsupervised WSD



Some (naïve) assumptions

Annotator inspects target word in contexts

Matches tokens to dictionary entries (using native intuition)

That’s all there is to it (?)

It’s not so simple!

- Comparison of multiple dictionaries shows little agreement:
- Lexicographers (speakers) carve up semantic space occupied by a word in different ways
- Different assumptions of “related senses”
- Fine-grained vs. coarse-grained distinctions (splitters vs. lumpers)
- Missing senses (deliberate omission of specialized uses of words; non-standard usages, new words, new meanings)

It’s not so simple!

Dictionaries were not made for annotation and word sense disambiguation

Made for look-up of unknown words, senses

User stops look-up when the unknown word/sense that she encountered is explained

User doesn’t need to examine all senses

(Kilgarriff 1990)

Overlap, duplicate senses are unproblematic for lexicography…

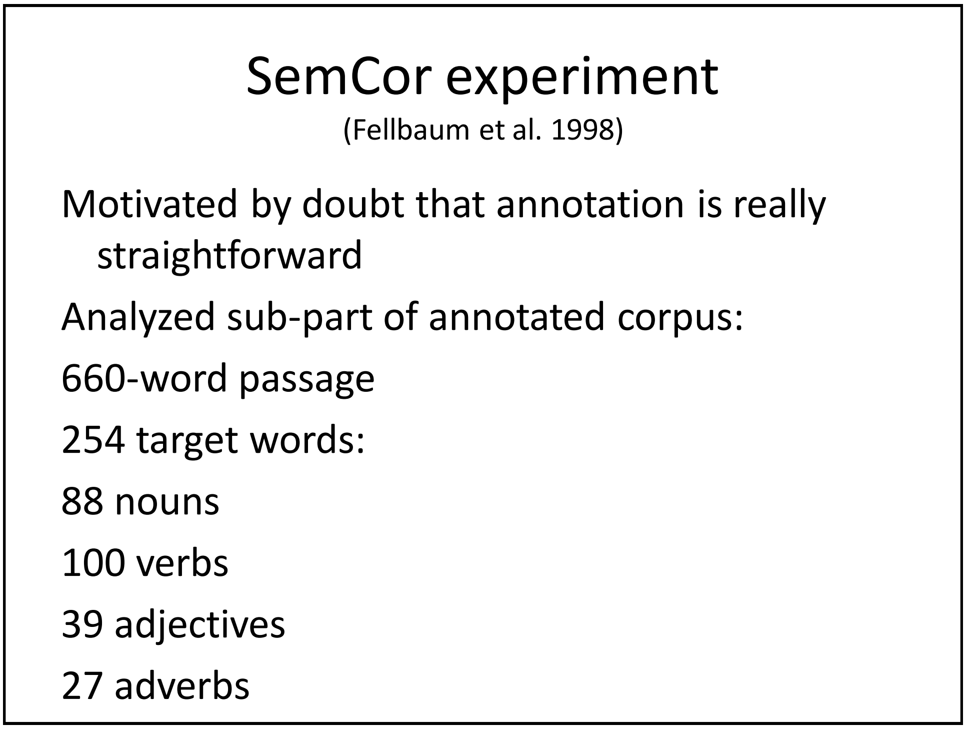
…but problematic for annotation!

Three experiments with manual semantic annotation

- SemCor
- WordNet gloss corpus
- MASC

Semantic Concordance (SemCor)

- First semantically annotated corpus (mid-1990s)
- parts of Brown Corpus
- novel *Red Badge of Courage*
- sequential, not targeted, annotation



SemCor experiment

(Fellbaum et al. 1998)

Motivated by doubt that annotation is really straightforward

Analyzed sub-part of annotated corpus:

660-word passage

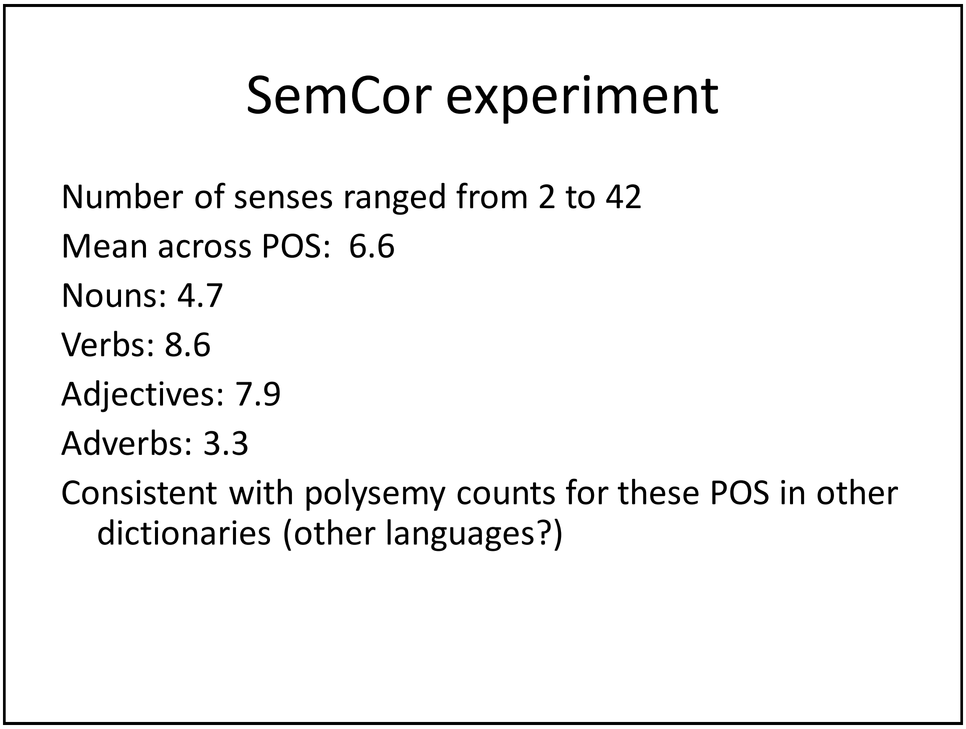
254 target words:

88 nouns

100 verbs

39 adjectives

27 adverbs



SemCor experiment

Number of senses ranged from 2 to 42

Mean across POS: 6.6

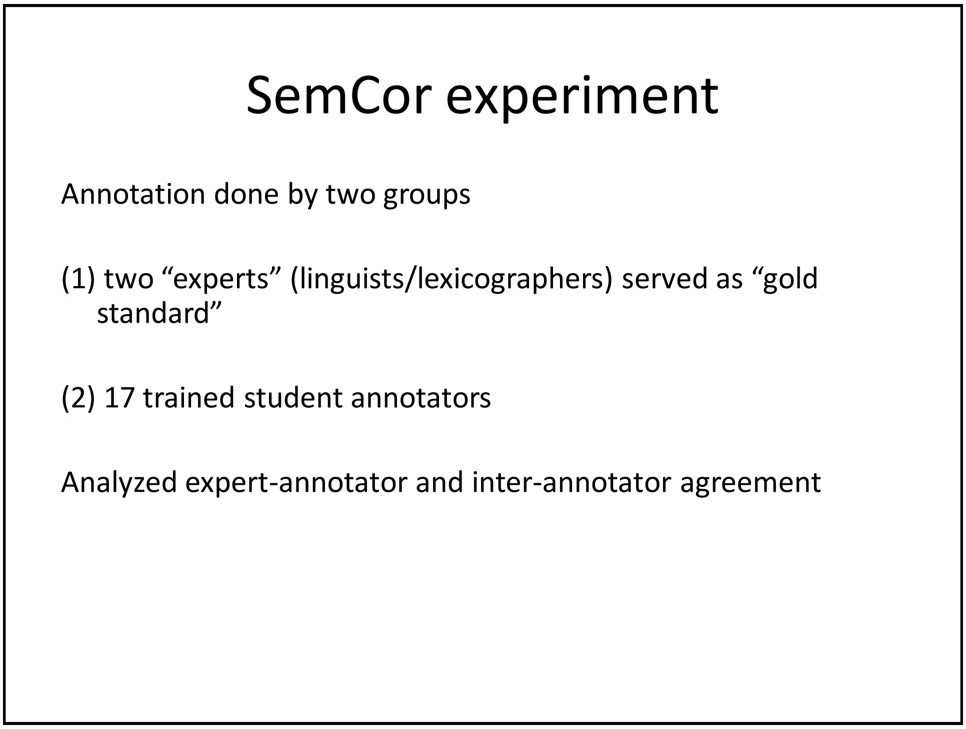
Nouns: 4.7

Verbs: 8.6

Adjectives: 7.9

Adverbs: 3.3

Consistent with polysemy counts for these POS in other dictionaries (other languages?)



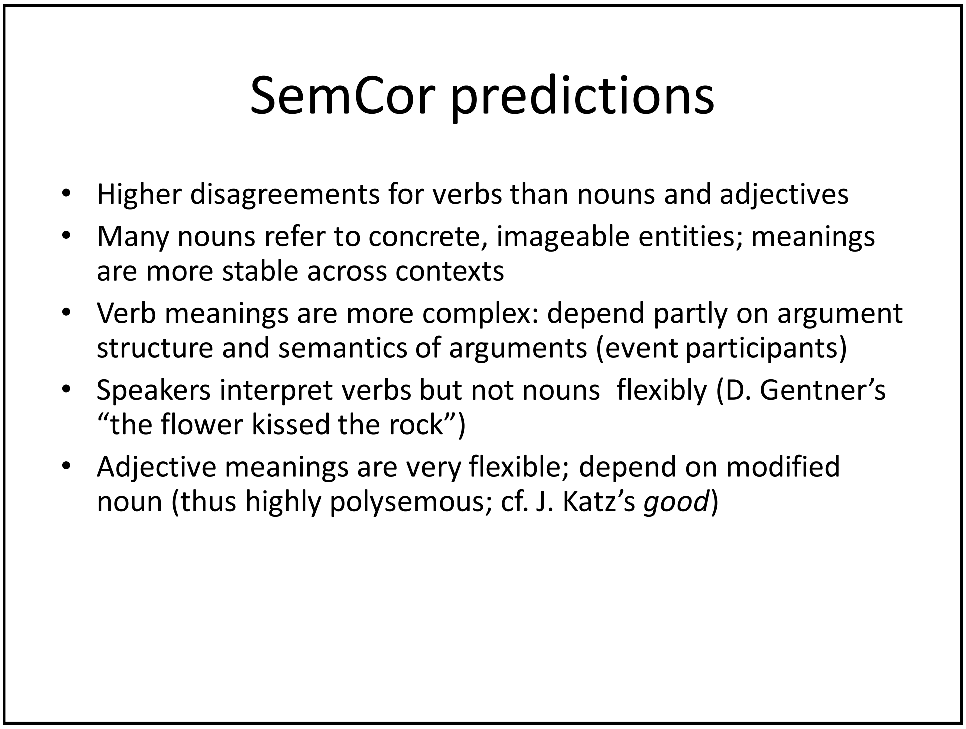
SemCor experiment

Annotation done by two groups

(1) two “experts” (linguists/lexicographers) served as “gold standard”

(2) 17 trained student annotators

Analyzed expert-annotator and inter-annotator agreement



SemCor predictions

- Higher disagreements for verbs than nouns and adjectives
- Many nouns refer to concrete, imageable entities; meanings are more stable across contexts
- Verb meanings are more complex: depend partly on argument structure and semantics of arguments (event participants)
- Speakers interpret verbs but not nouns flexibly (D. Gentner’s “the flower kissed the rock”)
- Adjective meanings are very flexible; depend on modified noun (thus highly polysemous; cf. J. Katz’s *good*)

SemCor predictions

Disagreement rate increases with number of senses (polysemy, not homonymy)

SemCor experiment

WordNet sense inventory was presented in two conditions

(1) Frequency order (previously annotated senses)

(2) Randomly scrambled order

SemCor predictions

First (most frequently tagged) sense is usually the most salient, broadest

Annotators prefer it

May save examining remaining senses?

Results (Overview)

- Overall agreement of annotators with “experts” was 72%
- Overall inter-annotator agreement was 82%
- Sharp drop-off in agreement with increasing sense number (polysemy)
- Significantly higher agreement rate for first sense
- Higher agreement for nouns than for verbs

Results (Overview)

Annotators were asked to rate **confidence** with which they chose senses

Overall high (1.8 on a scale 1-5)

Lower confidence for verbs than for nouns, for highly polysemous words

Higher confidence for random sense order (confirms that first-sense choice was not available as a shortcut)

Lessons learned

Sense annotation (word sense disambiguation) is feasible but hard

Difficulty depends on POS, degree of polysemy

Strong preference for broader/frequent sense

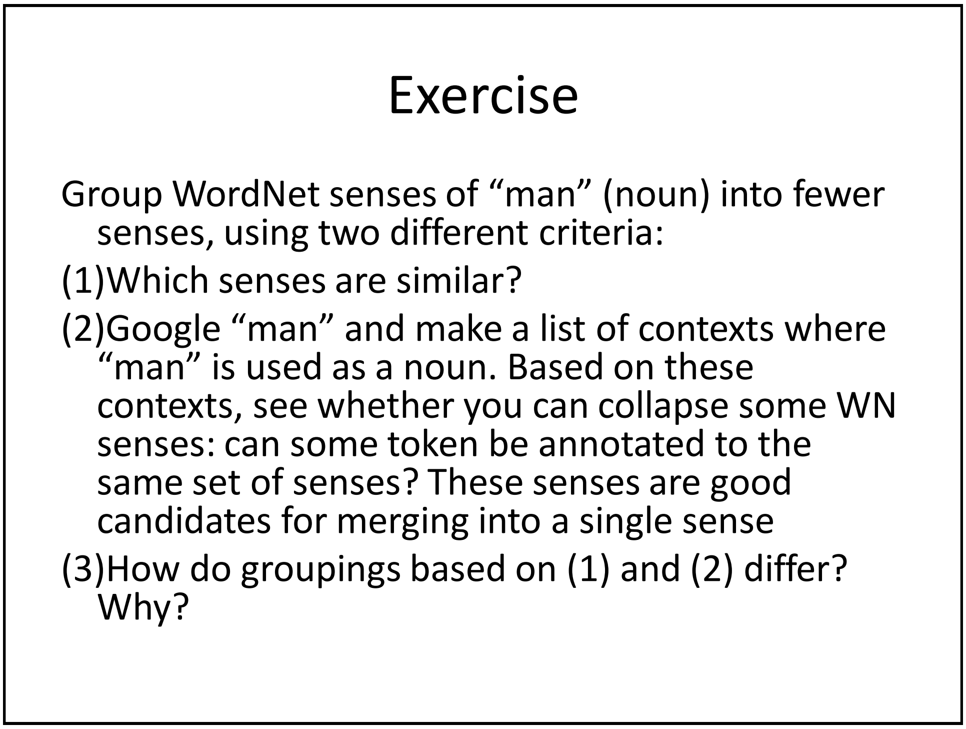
“expert”-”naïve” annotator difference—not sure what to make of that (but group size differed significantly)

Agreement rates found in SemCor experiment are not good enough for NLP appliations

Not better than “dumb” most frequent sense choice

What we can do

- Reduce sense inventory? (Note that WordNet’s inventory is *smaller* than that of standard dictionaries)
- Grouping into supersenses (underspecified) and subsenses (to accommodate context-specific readings)
- But: there are many grouping criteria, some conflicting (semantic, syntactic, domain,...)



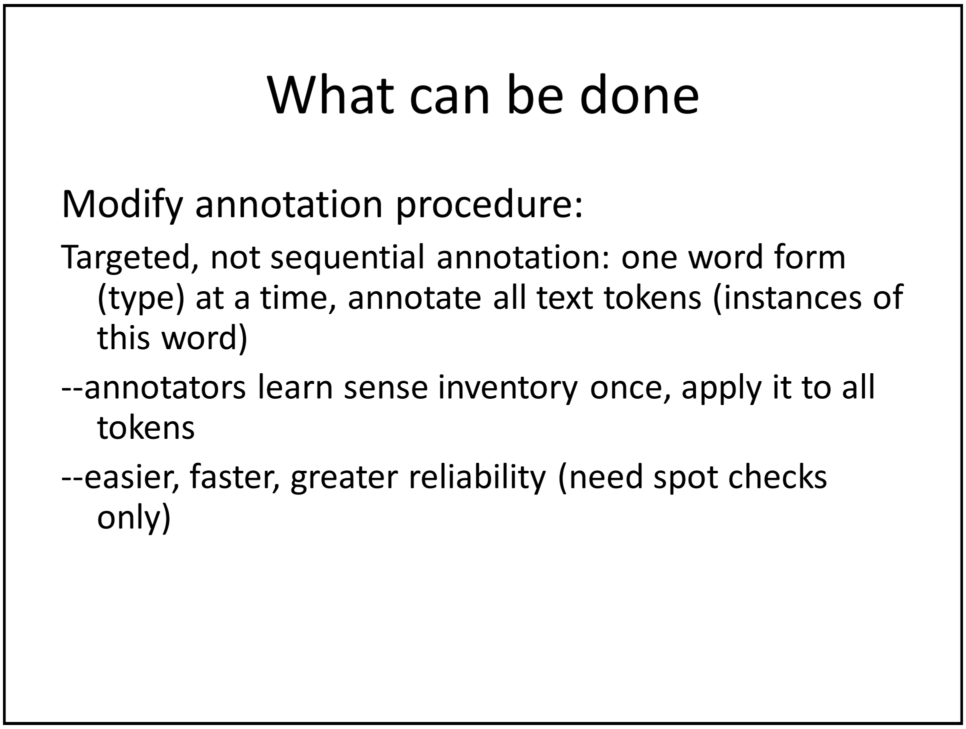
Exercise

Group WordNet senses of “man” (noun) into fewer senses, using two different criteria:

(1)Which senses are similar?

(2)Google “man” and make a list of contexts where “man” is used as a noun. Based on these contexts, see whether you can collapse some WN senses: can some token be annotated to the same set of senses? These senses are good candidates for merging into a single sense

(3)How do groupings based on (1) and (2) differ? Why?



What can be done

Modify annotation procedure:

Targeted, not sequential annotation: one word form (type) at a time, annotate all text tokens (instances of this word)

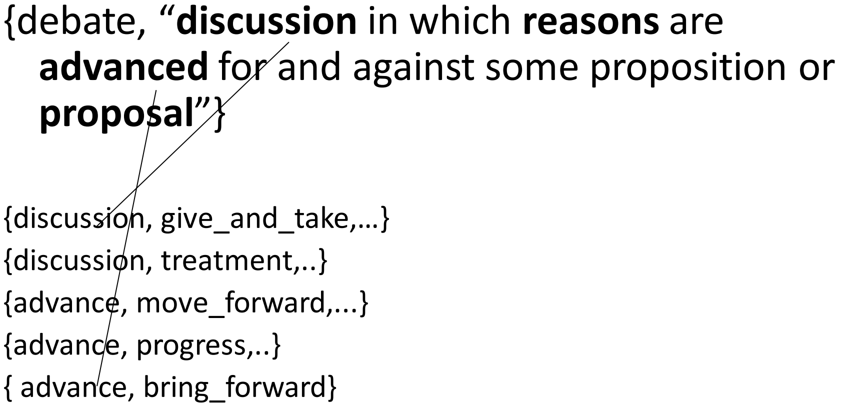
--annotators learn sense inventory once, apply it to all tokens

--easier, faster, greater reliability (need spot checks only)

Second try: WordNet Gloss corpus

- Glosses: definitions in WordNet’s sense entries (synsets)
- Annotate nouns, verbs, adjectives in glosses against WN synsets
- Closed system: annotated glosses constitute a corpus; for a given sense, glosses provide contexts
- http://wordnet.princeton.edu/wordnet/download/

Gloss (definition) Annotation



{debate, “**discussion** in which **reasons** are **advanced** for and against some proposition or **proposal**”}

{discussion, give_and_take,…}
{discussion, treatment,..}
{advance, move_forward,...}
{advance, progress,..}
{ advance, bring_forward}

Gloss Tagging

Most words are monosemous and can be tagged automatically (monosemy is relative to WordNet—potentially flawed!)

Metalinguistic words in the glosses are not tagged:

“to scowl is to grimace in some **manner**”

Manner is not meaningfully related to *scowl*,

but *grimace* is

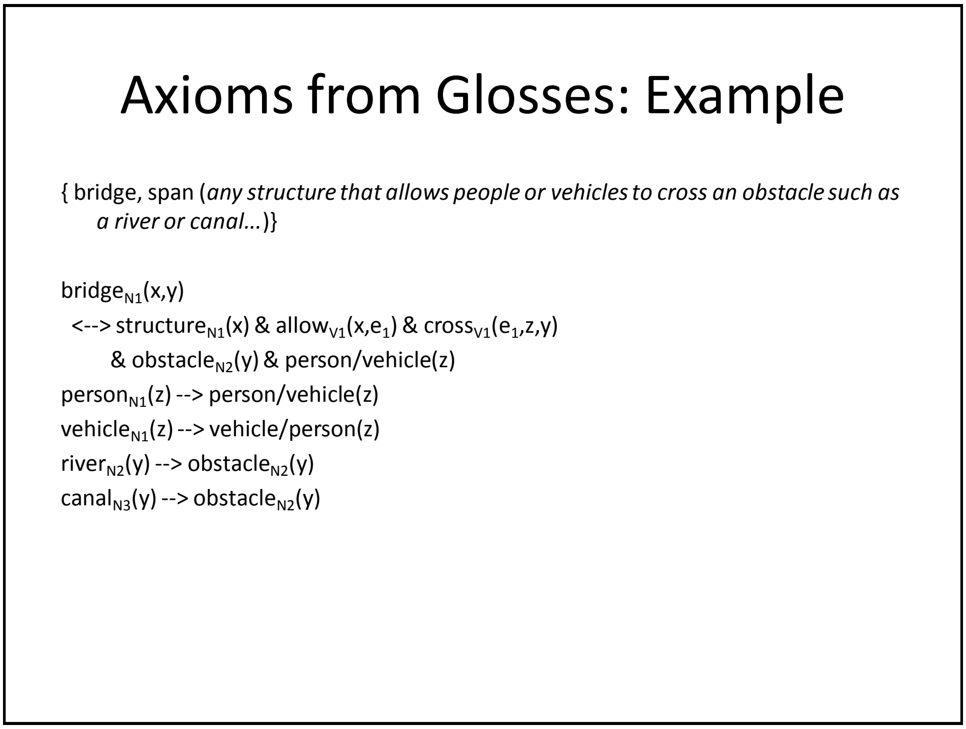
So only *grimace* can provide useful information for ML!

Gloss corpus

Glosses were translated into Logical Form (Hobbs)

Variables were indexed with WordNet senses

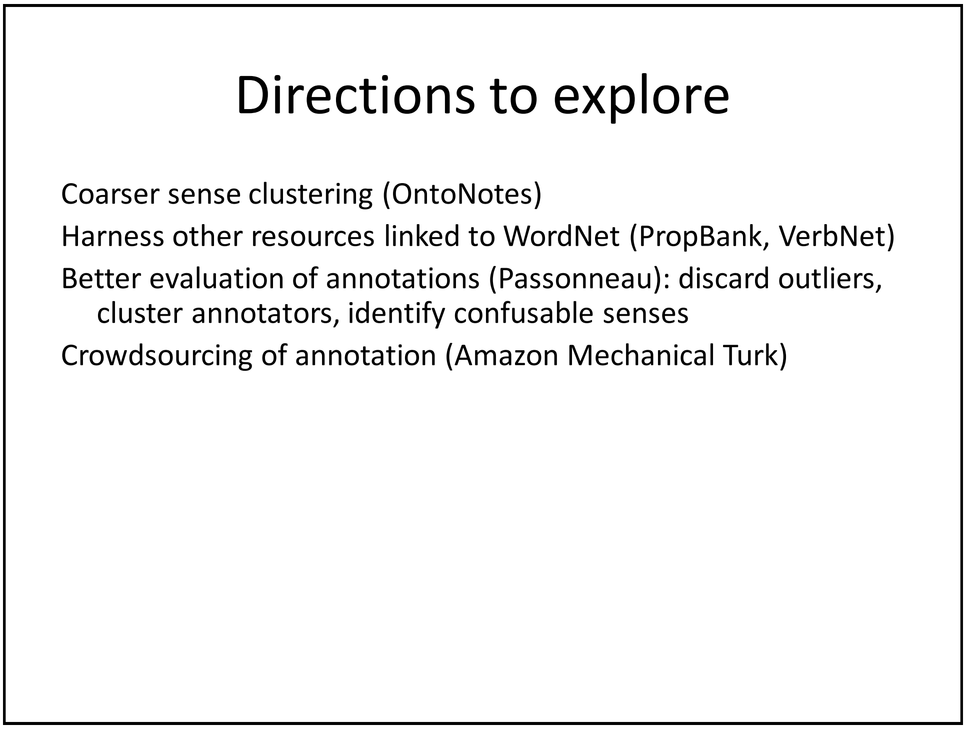
Goal: provide automatic systems with reasoning capabilities (inferencing, recognizing entailments)



Axioms from Glosses: Example

{ bridge, span (*any structure that allows people or vehicles to cross an obstacle such as a river or canal…*)}

bridge$_{N1}$(x,y)
<--> structure$_{N1}$(x) & allow$_{V1}$(x,e$_1$) & cross$_{V1}$(e$_1$,z,y)
& obstacle$_{N2}$(y) & person/vehicle(z)
person$_{N1}$(z) --> person/vehicle(z)
vehicle$_{N1}$(z) --> vehicle/person(z)
river$_{N2}$(y) --> obstacle$_{N2}$(y)
canal$_{N3}$(y) --> obstacle$_{N2}$(y)



Directions to explore

Coarser sense clustering (OntoNotes)
Harness other resources linked to WordNet (PropBank, VerbNet)
Better evaluation of annotations (Passonneau): discard outliers, cluster annotators, identify confusable senses
Crowdsourcing of annotation (Amazon Mechanical Turk)

Intermediate summary

Manual tagging is expensive, slow

Often not much more reliable than “first/most frequent-sense” rule

Large semantically annotated corpora are still elusive

Scale up with (semi-)automatic annotation, using WordNet relations (to be continued)

WordNet: a resource for bridging the lexical bottleneck in NLP

Christiane Fellbaum

Outline

- What is WordNet and why is it interesting/useful?
- A bit of history
- WordNet for natural language processing/word sense disambiguation
- Using “evocation” to augment WordNet
- Multilingual WordNets
- WordNet for multi-modal information processing

What is WordNet?

- A large lexical database, semantic resource, “electronic dictionary,” developed and maintained at Princeton University

  http://wordnet.princeton.edu
- Includes most English nouns, verbs, adjectives, adverbs
- Electronic format makes it accessible and useful for automatic systems
- Used in many Natural Language Processing applications requiring semantic analysis (information retrieval, text mining, question answering, machine translation, AI/reasoning,...)

What’s special about WordNet?

- Traditional paper dictionaries are organized alphabetically
- As a result, words that are found together (on the same page) are not related by *meaning*
- WordNet is organized by meaning: words in close proximity are semantically similar
- Human users and computers can browse WordNet and find words that are meaningfully related to their queries (somewhat like in a hyperdimensional thesaurus)
- Meaning similiarity can be measured and quantified to support Natural Language Understanding, in particular WSD

A bit of history

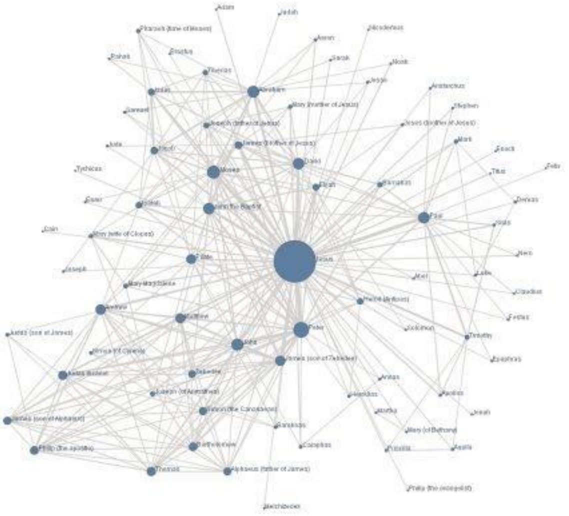
Late 1960s, 70s: Artificial Intelligence (AI)

How do humans store and access knowledge about concept?

Hypothesis: concepts are interconnected via meaningful relations

Semantic network representation

(Collins and Quillian 1969, 1970, 1972)

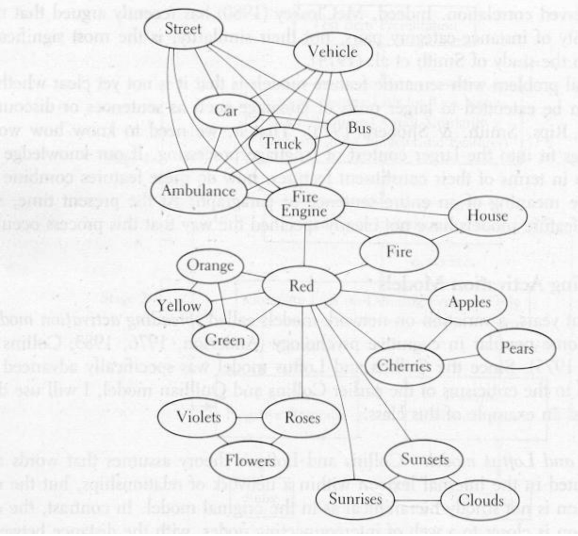


Theory of semantic processing

Spreading Activation (Collins and Loftus, 1975)

A node in the network (a concept/word) gets activated and activates other, nearby nodes

Links among nodes are weighted



What would such a network look like exactly?

Assumptions

Knowledge of concepts

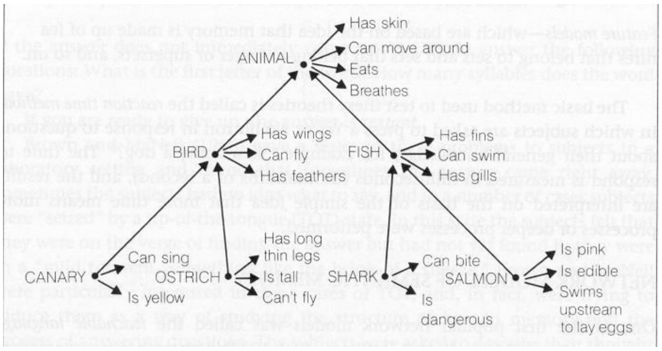
--stored economically in our minds/brains

--computed "on the fly"

--via access to general concepts

E.g., we know that "canaries fly" because "birds fly" and "canaries are a kind of bird"

Collins & Quillian Semantic network



A model of semantic organization

Knowledge is stored **only once** at the highest possible node and inherited downward (not re-stored)

animals move, birds fly, canaries sing

no redundant storage: birds move, canaries fly

no upward inheritance: *animals fly and sing

Collins & Quillian (1969) measured reaction times to statements involving knowledge distributed across different "levels"

Collins & Quillian experiment

Responses to statements like

Do birds move?

Do canaries move?

Do canaries have feathers?

Are canaries yellow?

Reaction times varied depending on how many nodes had to be traversed to access the information

Critique

Results are not compelling

reaction times are influenced (at least) by

--prototypicality (how typical an exemplar of the category bird is canary?)

--word frequency (statement with robin might be processed faster than with canary)

--category size (how many birds and associated information has to be searched/discarded?)

--uneven semantic distance across levels (big jump from animal to bird; smaller jump from canary to bird)

But the idea inspired WordNet (1986), which asked:

Can most/all of the lexicon (of any language?) be represented as a semantic network?

Would some words be left hanging in space? (If so, which ones?)

WordNet

If the (English) lexicon can be represented as a semantic network (a graph), what are the links that connect the nodes?

Links among nodes (concepts) are conceptual-semantic

Links among specific words are lexical

Lexical links subsume conceptual-semantic links

Whence the relations?

Inspection of association norms:
stimulus: *hand* reponse: *finger, arm*
stimulus: *help* response: *aid*
stimulus: *thin* response: *fat*
stimulus: *rodent* response: *rat*

Speech errors: substitution of, e.g., *week* for *day*

Such data show systematic relations among words
Lexicon-as-library metaphor

To be continued