Word Sense Disambiguation

Bill MacCartney

CS224U

17 January 2012

Lexical ambiguity

- The meaning of *bass* depends on context
  - Are we talking about music, or fish?

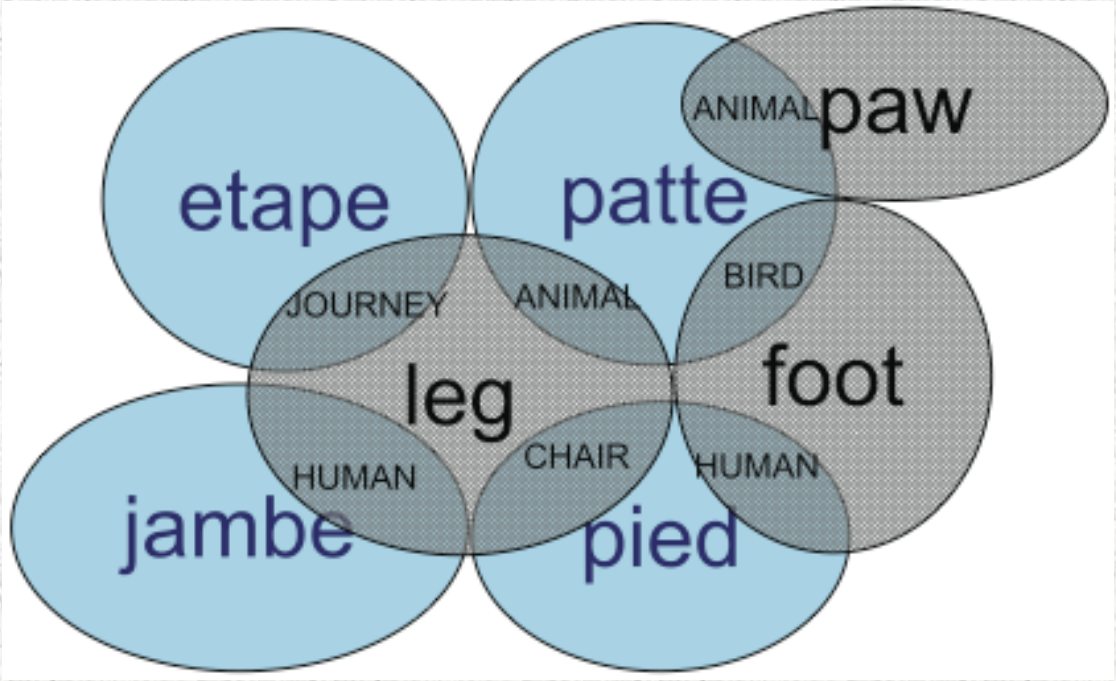
    An electric guitar and bass player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.

    And it all started when fishermen decided the striped bass in Lake Mead were too skinny.
- These senses translate differently into other languages

WordNet Sense	Spanish Translation	Roget Category	Target Word in Context
bass^4	lubina	FISH/INSECT	...fish as Pacific salmon and striped **bass** and...
bass^4	lubina	FISH/INSECT	...produce filets of smoked **bass** or sturgeon...
bass^7	bajo	MUSIC	...exciting jazz **bass** player since Ray Brown...
bass^7	bajo	MUSIC	...play **bass** because he doesn't have to solo...

Figure 20.1 Possible definitions for the inventory of sense tags for *bass*.

Lexical ambiguity across languages



Hutchins & Somers 1992

Homonymy & polysemy

- In fact, *bass* has 8 senses in WordNet (as a noun)
- It is both homonymous and polysemous

Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- S: (n) **bass**, basso (an adult male singer with the lowest voice)
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Ambiguity is rampant!

I saw a man who is 98 years old and can still walk and tell jokes

26 senses	11 senses	4 senses	8 senses	5 senses	4 senses	10 senses	8 senses	3 senses

43,929,600 senses

The WSD task

- The Word Sense Disambiguation (WSD) task
  - To identify the intended sense of a word *in context*
  - Usually assumes a fixed inventory of senses (e.g., WordNet)
- Can be viewed as categorization / tagging task
  - So, similar to the POS tagging task
  - But, there are important differences! → upper bound is lower
- Differs from Word Sense *Discrimination* task
  - Clustering usages of a word into different senses, without regard to any particular sense inventory. Uses unsupervised techniques.
- WSD is crucial prerequisite for many NLP applications (?)
  - WSD is not itself an end application
  - But many other tasks seem to require WSD (examples?)
  - In practice, the implementation path hasn't always been clear

WSD task variants

- *Lexical sample task*: WSD for small, fixed set of words
  - E.g. *line*, *interest*, *plant*
  - Focus of early work in WSD
  - Supervised learning works well here
- *All-words task*: WSD for every content word in a text
  - Like POS tagging, but much larger tag set (varies by word)
  - Big data sparsity problem — don't have labeled data for every word!
  - Can't train separate classifier for every word
- SENSEVAL includes both tasks

Early days of WSD

- Noted as a problem for machine translation (Weaver, 1949)
  - E.g., a *bill* in English could be a *pico* or a *cuenta* in Spanish
  - One of the oldest problems in NLP!
- Bar-Hillel (1960) posed the following problem:
  - *Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.*
  - Is "pen" a writing instrument or an enclosure where children play?
- …declared it unsolvable, and left the field of MT (!):

  "Assume, for simplicity's sake, that *pen* in English has only the following two meanings: (1) a certain writing utensil, (2) an enclosure where small children can play. I now claim that no existing or imaginable program will enable an electronic computer to determine that the word *pen* in the given sentence within the given context has the second of the above meanings, whereas every reader with a sufficient knowledge of English will do this 'automatically'." (1960, p. 159)

Changing approaches to WSD

- Early WSD work: semantic networks, frames, logical reasoning, expert systems
  - However, the problem got quite out of hand
  - The word expert for *throw* is "currently six pages long, but should be ten times that size" (Small & Rieger 1982)
- Supervised machine learning & contextual features
  - Great success, beginning in early 90s (Gale et al. 92)
  - But, requires expensive hand-labeled training data
- Search for ways to minimize need for hand-labeled data
  - Dictionary- and thesaurus-based approaches (e.g., Lesk)
  - Semi-supervised approaches (e.g., Yarowsky 95)
  - Leveraging parallel corpora, web, Wikipedia, etc. (e.g., Mihalcea 07)

Supervised WSD

- Start with sense-annotated training data
- Extract features describing contexts of target word
- Train a classifier using some machine learning algorithm
- Apply classifier to unlabeled data

- WSD was an early paradigm of applying supervised machine learning to NLP tasks!

WSD Corpora

- Supervised approach requires sense-annotated corpora
  - Hand-tagging of senses can be laborious, expensive, unreliable
  - Unannotated data can also be useful: newswire, web, Wikipedia
- Sense-annotated corpora for lexical sample task
  - *line-hard-serve* corpus (4000 examples)
  - *interest* corpus (2400 examples)
  - SENSEVAL corpora (with 34, 73, and 57 target words, respectively)
  - DSO: 192K sentences from Brown & WSJ (121 nouns, 70 verbs)
- Sense-annotated corpora for all-words task
  - SemCor: 200K words from Brown corpus w/ WordNet senses
    - SemCor frequencies determine *ordering* of WordNet senses
  - SENSEVAL 3: 2081 tagged content words

SENSEVAL data: *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.

SemCor data

```
<contextfile concordance="brown">
<context filename="br-h15" paras="yes">
…..
<wf cmd="ignore" pos="IN">in</wf>
<wf cmd="done" pos="NN" lemma="fig" wnsn="1" lexsn="1:10:00::">fig.</wf>
<wf cmd="done" pos="NN" lemma="6" wnsn="1“ lexsn="1:23:00::">6</wf>
<punc>)</punc>
<wf cmd="done" pos="VBP" ot="notag">are</wf>
<wf cmd="done" pos="VB" lemma="slip" wnsn="3" lexsn="2:38:00::">slipped</wf>
<wf cmd="ignore" pos="IN">into</wf>
<wf cmd="done" pos="NN" lemma="place" wnsn="9" lexsn="1:15:05::">place</wf>
<wf cmd="ignore" pos="IN">across</wf>
<wf cmd="ignore" pos="DT">the</wf>
<wf cmd="done" pos="NN" lemma="roof" wnsn="1" lexsn="1:06:00::">roof</wf>
<wf cmd="done" pos="NN" lemma="beam" wnsn="2" lexsn="1:06:00::">beams</wf>
<punc>,</punc>
```

Features for supervised WSD

- Features should describe *context* of target word
  - "You shall know a word by the company it keeps" — Firth 1957
- Preprocessing of target sentence
  - POS tagging, lemmatization, syntactic parsing?
- Collocational features: specific positions relative to target
  - E.g., words at index –3, –2, –1, +1, +2, +3 relative to target
  - Features typically include word identity, word lemma, POS
- Bag-of-words features: general neighborhood of target
  - Words in symmetric window around target, ignoring position
  - Binary word occurrence features (so, actually set-of-words)
  - Often limited to words which are frequent in such contexts

Feature extraction example

An electric guitar and bass player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.

Collocational features	
word_L3	electric
POS_L3	JJ
word_L2	guitar
POS_L2	NN
word_L1	and
POS_L1	CC
word_R1	player
POS_R1	NN
word_R2	stand
POS_R2	VB
word_R3	off
POS_R3	RB

Bag-of-words features	
fishing	0
big	0
sound	0
player	1
fly	0
rod	0
pound	0
double	0
runs	0
playing	0
guitar	1
band	0

Naïve Bayes for WSD

- A Naïve Bayes classifier chooses the most likely sense for a word given the features of the context:

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} \; P(s \mid \vec{f})$$

- Using Bayes' Law, this can be expressed as:

$$\hat{s} = \underset{s \in S}{\arg\max} \frac{P(s)P(\vec{f} \mid s)}{P(\vec{f})} = \underset{s \in S}{\arg\max} P(s)P(\vec{f} \mid s)$$

- The "naïve" assumption: all the features are conditionally independent, given the sense:

$$\hat{s} = \underset{s \in S}{\arg\max} P(s) \prod_{j=1}^{n} P(f_j \mid s)$$

Naïve Bayes training

- Set parameters of Naïve Bayes using maximum likelihood estimation (MLE) from training data
- In other words, just count!

$$P(s_i) = \frac{count(s_i, w_j)}{count(w_j)} \qquad P(f_j \mid s) = \frac{count(f_j, s)}{count(s)}$$

- Naïve Bayes is dead-simple to implement, but …
  - Numeric underflow → use log probabilities
  - Zero probabilities → use smoothing

Gale, Church, & Yarowsky 92

- Used Naïve Bayes to disambiguate six polysemous nouns
  - *duty, drug, land, language, position, sentence*
- Used an *aligned corpus* (Hansard) to get the word senses

English	French	Sense	# examples
duty	droit devoir	tax obligation	1114 691
drug	medicament drogue	medical illicit	2292 855
land	terre pays	property country	1022 386

- Bag-of-words features: what words appear in context?

Gale et al. 92: Results

- Achieved ~90% accuracy — seems very good!
  - But, it was a binary decision problem
  - Also, you're choosing between quite different senses
  - Of course, that may be the most important case to get right…
- Good context clues for *drug*:
  - medication: *prices, prescription, patent, increase*
  - illegal substance: *abuse, paraphernalia, illicit, alcohol, cocaine, traffickers*
- Also evaluated impact of changing context window size …

Gale et al.: remote context is informative

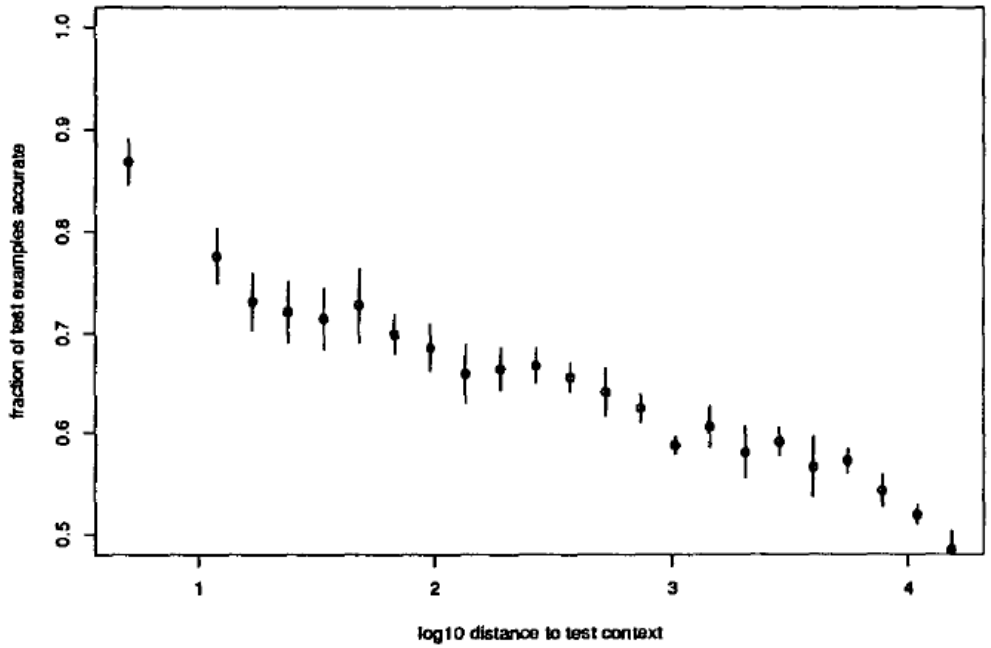


Figure II. Remote context is informative.
The horizontal axis shows the distance of context words from an ambiguous word, while the vertical scale shows the percent correct when using ten context words at the specified distance in doing the disambiguation. The vertical lines show the mean and standard deviation of mean for six disambiguations. With two equiprobable choices, 50 percent represents chance performance. Performance remains above chance for ten word contexts up to ten thousand words away from the ambiguous word.

Gale et al.: wide contexts are useful

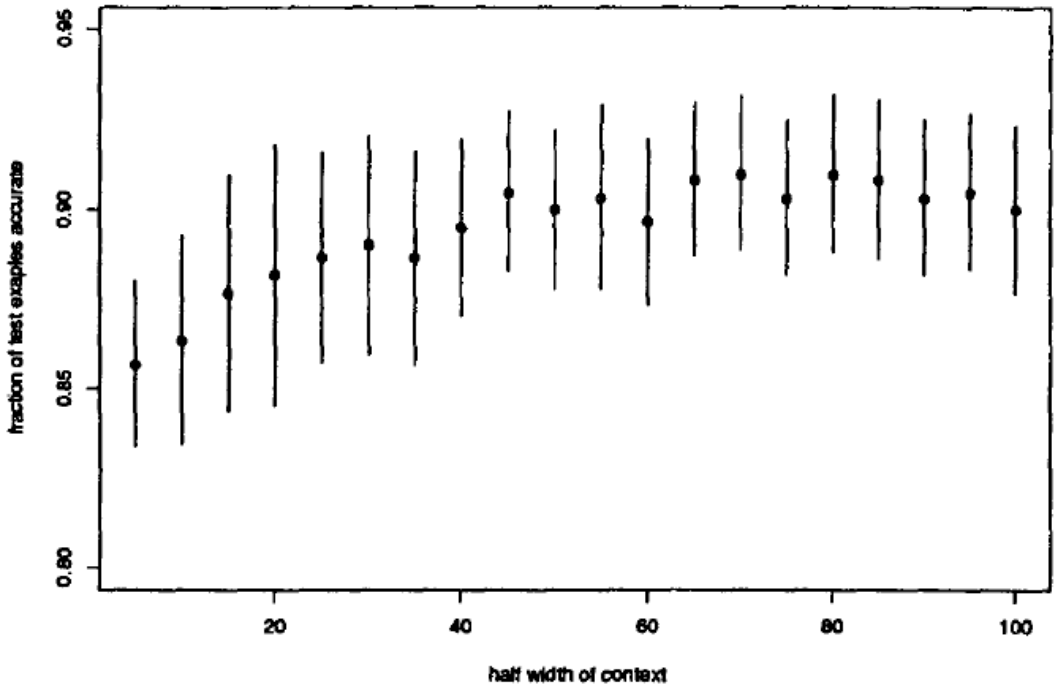


Figure III. Wide contexts are useful.
The horizontal axis shows the maximum distance of context words from an ambiguous word, while the vertical scale shows the percent correct when using all context words out to the specified distance in disambiguation. While performance rises very rapidly with the first few words, it clearly continues to improve through about twenty words, and is not worse by fifty words.

Gale et al.: Learning curve

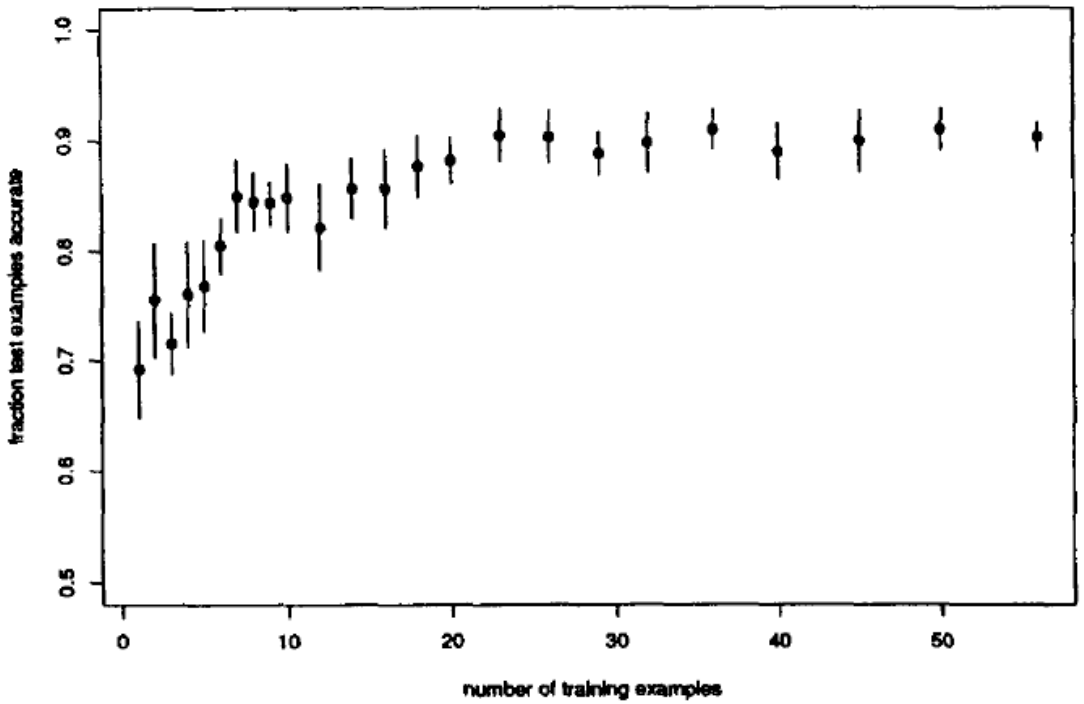


Figure IV. Just a few training examples do surprisingly well.
The horizontal axis shows the number of examples used in training while the vertical scale shows the mean percent correct in six disambiguations. The performance increases rapidly for the first few examples, and seems to have reached a maximum by 50 or 60 examples.

Decision list classifiers for WSD

- A sequence of tests on features of context
  - Analogous to a `case` statement in programming
  - Each case yields a particular sense prediction if matched
  - Default case: most frequent sense
  - Tests can consider both collocational & bag-of-words features

Rule		Sense
fish within window	⇒	**bass**^{1}
striped bass	⇒	**bass**^{1}
guitar within window	⇒	**bass**^{2}
bass player	⇒	**bass**^{2}
piano within window	⇒	**bass**^{2}
tenor within window	⇒	**bass**^{2}
sea bass	⇒	**bass**^{1}
play/V *bass*	⇒	**bass**^{2}
river within window	⇒	**bass**^{1}
violin within window	⇒	**bass**^{2}
salmon within window	⇒	**bass**^{1}
on bass	⇒	**bass**^{2}
bass are	⇒	**bass**^{1}

Learning a decision list classifier

- How to learn a decision list classifier?
- Yarowsky 94 proposes a method for binary WSD:
  - consider all feature-value pairs
  - order them by log-likelihood ratio

$$\left|\log\left(\frac{P(Sense_1|f_i)}{P(Sense_2|f_i)}\right)\right|$$

- (Quite different from standard decision list learning)

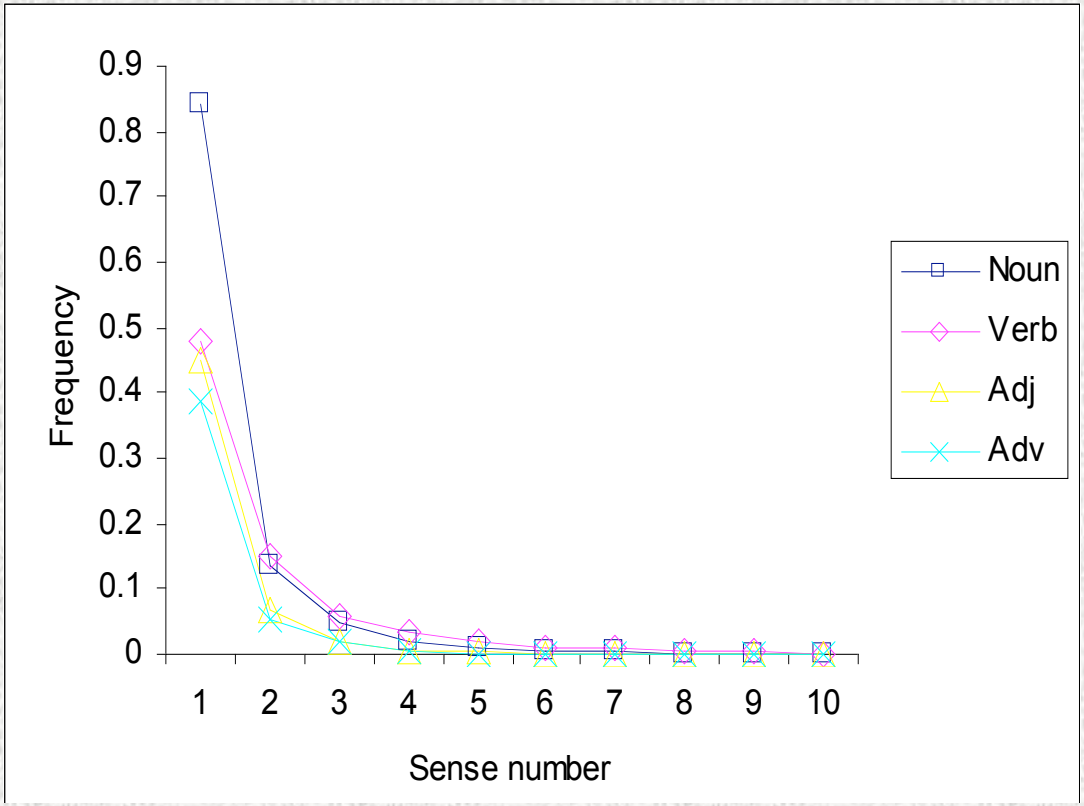
Evaluation of WSD systems

- Extrinsic (task-based, end-to-end, *in vivo*) evaluation
  - evaluate MT, IR, QA, … system with and without WSD system
  - only way to tell whether WSD is helping on some real application
  - but: difficult, time-consuming, may not generalize to other apps
- Intrinsic (*in vitro*) evaluation
  - apply WSD system to hand-labeled test data (e.g., SemCor, SENSEVAL)
  - measure accuracy (or P/R/F1) in matching gold-standard labels
- Need baseline evaluation, for comparison
  - Random is weak: 14% accuracy on SENSEVAL-2 lexical sample task
  - Stronger baselines: most-frequent sense (MFS), Corpus Lesk (below)
- Also need ceiling: human inter-annotator agreement
  - typically 75-80% for all-words task using WordNet-style senses
  - up to 90% for more coarse-grained (or binary) sense inventories

The MFS baseline

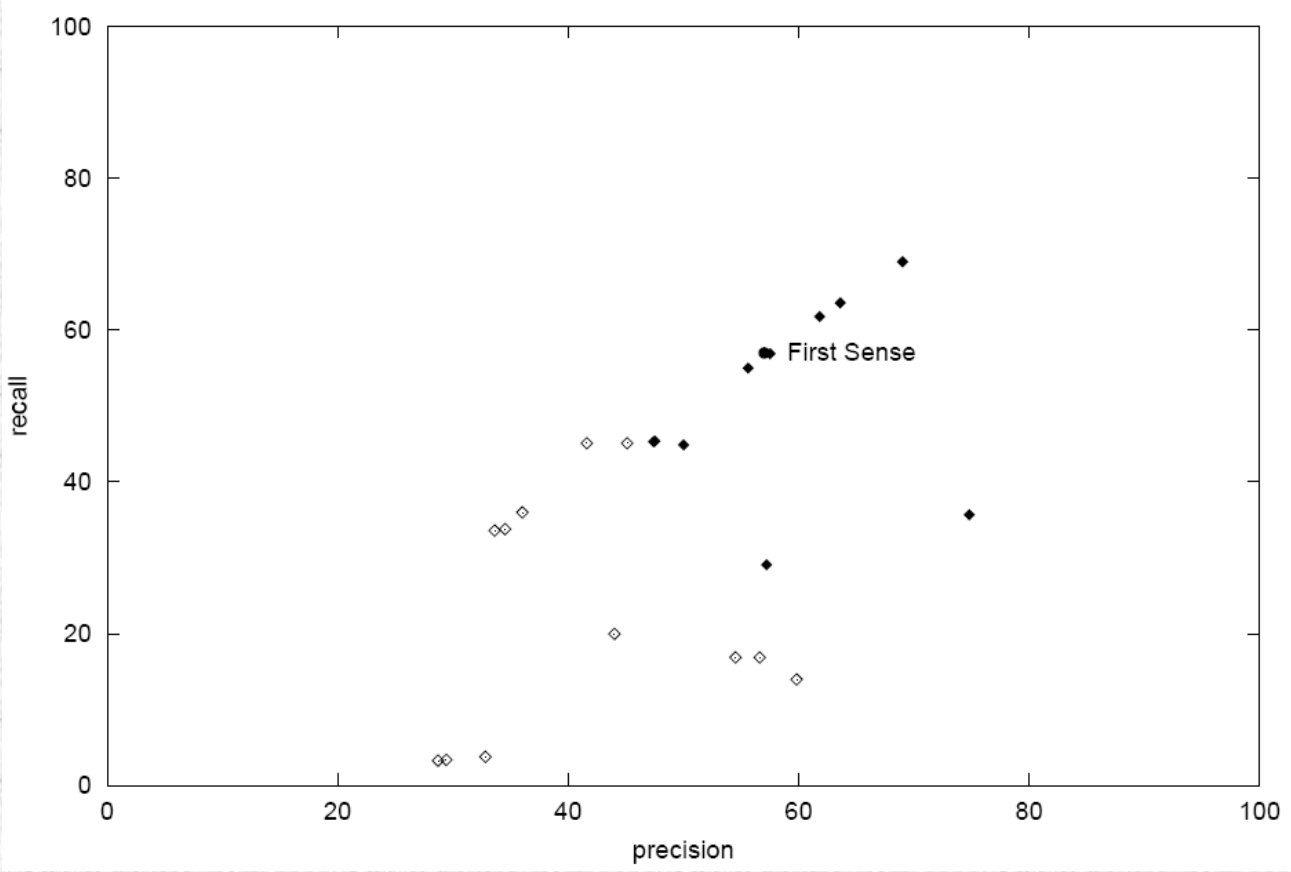
- predict most frequent sense (MFS) in some labeled corpus
  - MFS in SemCor → first WordNet sense
- a surprisingly strong baseline
  - often 50-60% accuracy on lexical sample task w/ WordNet senses
  - even higher with coarser senses, more skewed distributions
  - often tough to beat, esp. on all-words task
- problem: doesn't take account of context / genre
  - MFS of *star* in SemCor is *celestial body*
  - but for WSD on popular news, *celebrity* would be preferred
- problem: subject to quirks of corpus, sparsity
  - *tiger* rare in SemCor: first sense in WordNet is *audacious person*
  - *embryo* not in SemCor: 1st in WN is *rudimentary plant*, not *fertilized egg*

Sense distributions are Zipfian



Sense distributions in SemCor

The power of the MFS heuristic



Performance of the MFS heuristic compared with systems in the SENSEVAL-2 English all-words task

Working around the data problem

- Supervised WSD methods yield best performance, but:
  - Training data is expensive to generate
  - Doesn't work for words not in training data
  - What about less-common languages (Catalan, Swahili, etc.)?
- Can we get indirect supervision?
  - Dictionary- and thesaurus-based approaches (e.g., Lesk)
  - Semi-supervised approaches (e.g., Yarowsky 95)
- Can we eschew supervision entirely?
  - Unsupervised approaches (e.g., Schütze 92, 98)
  - Word sense *discrimination* (clustering)
- Can we cleverly exploit other kinds of resources?
  - Leveraging parallel corpora, Wikipedia, etc. (e.g., Mihalcea 07)

Dictionary-based approaches

- Lesk (1986)
  - Retrieve all sense definitions of target word from MRD
  - Compare with sense definitions of words in context
  - Choose the sense with the most overlapping words
- Example
  - *pine*
    1. a kind of evergreen tree with needle-shaped leaves
    2. to waste away through sorrow or illness
  - *cone*
    1. A solid body which narrows to a point
    2. Something of this shape, whether solid or hollow
    3. Fruit of certain evergreen trees
  - Disambiguate: *pine cone*

Lesk variants

- Simplified Lesk
  - Retrieve all sense definitions of target word from MRD
  - Compare with ~~sense definitions of~~ words in context
  - Choose the sense with the most overlapping words
- Corpus Lesk
  - Include SEMCOR sentences in signature for each sense
  - Weight words by inverse document frequency (IDF)
  - IDF(w) = –log P(w)
  - Best-performing Lesk variant
  - Used as a (strong) baseline in SENSEVAL

Selectional Restrictions & Preferences

- Early knowledge source for WSD: *selectional restrictions*
  - *"In our house, everybody has a career and none of them includes washing dishes$_1$", he says.*
  - *Mrs. Chen works efficiently, stir-frying several simple dishes$_2$, including braised pig's ears and chicken livers …*
- Can we use this to disambiguate subjects? Verbs?
- Problem: selectional restrictions are often violated
  - *But it fell apart in 1931, perhaps because people realized that you can't eat gold for lunch if you're hungry*
- Solution: information-theoretic *selectional preferences*
- Resnik (1998): 44% accuracy using selectional preferences
  - OK for an unsupervised method, but worse than MFS or Lesk

Minimally supervised WSD

- The Yarowsky (1995) bootstrapping algorithm
  - start from small seed set of hand-labeled data Λ_0
  - learn decision-list classifier from Λ_0
  - use learned classifier to label unlabeled data V_0
  - move high-confidence examples in V_0 to Λ_1
  - repeat!
- Requires good confidence metric
  - Yarowsky used log-likelihood ratio of decision list rule that fired
- Can generate seed data using heuristics
  - One sense per collocation
    - Select informative collocates & extract examples from corpus
  - One sense per discourse
    - Validity depends on granularity of sense inventory

The Yarowsky algorithm

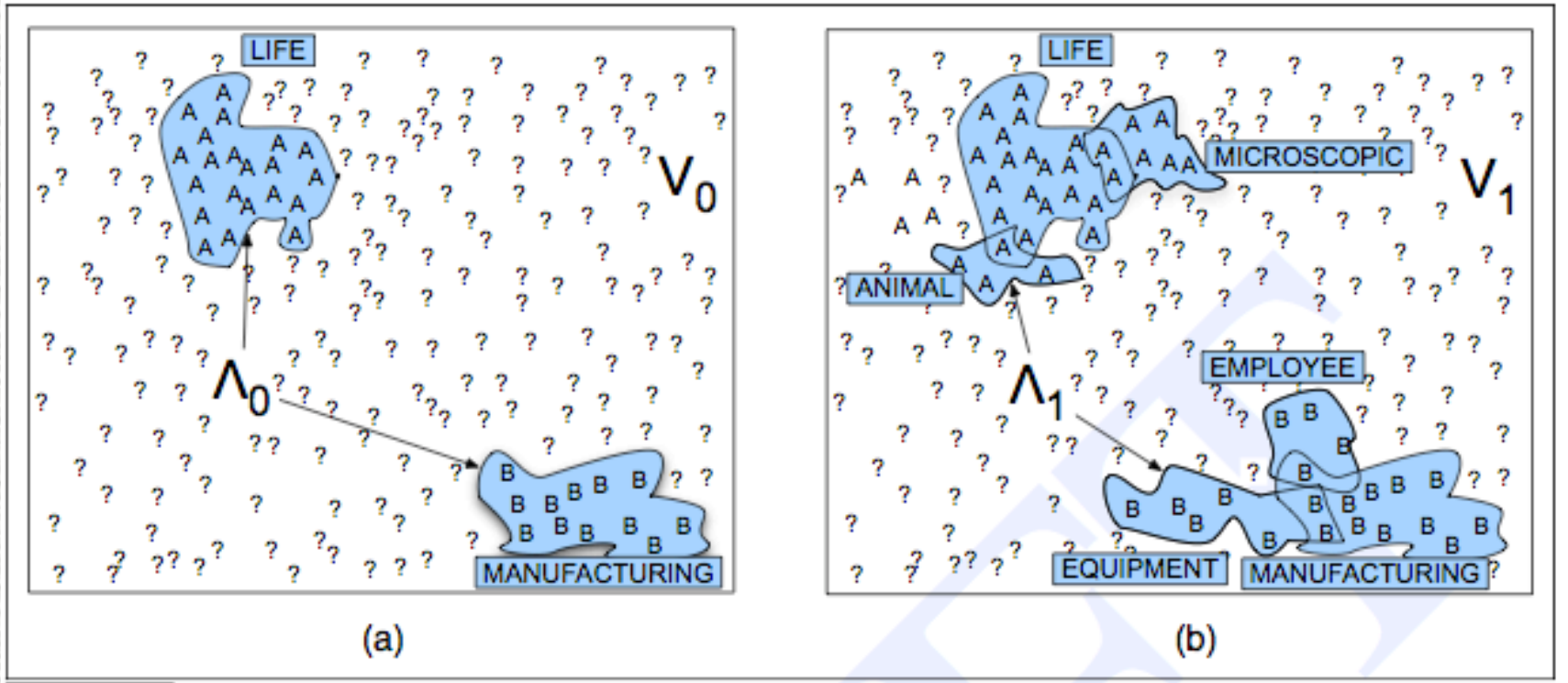


Figure 20.4 The Yarowsky algorithm disambiguating "plant" at two stages; "?" indicates an unlabeled observation, A and B are observations labeled as SENSE-A or SENSE-B. The initial stage (a) shows only seed sentences Λ_0 labeled by collocates ("life" and "manufacturing"). An intermediate stage is shown in (b) where more collocates have been discovered ("equipment", "microscopic", etc.) and more instances in V_0 have been moved into Λ_1, leaving a smaller unlabeled set V_1. Figure adapted from Yarowsky (1995).