Word Sense Disambiguation

Word Sense Disambiguation (WSD)

- Given
  - A word in context
  - A fixed inventory of potential word senses
  - Decide which sense of the word this is
- Why? Machine translation, QA, speech synthesis
- What set of senses?
  - English-to-Spanish MT: set of Spanish translations
  - Speech Synthesis: homographs like *bass* and *bow*
  - In general: the senses in a thesaurus like WordNet

Two variants of WSD task

- Lexical Sample task
  - Small pre-selected set of target words (*line, plant*)
  - And inventory of senses for each word
  - **Supervised machine learning: train a classifier for each word**
- All-words task
  - Every word in an entire text
  - A lexicon with senses for each word
  - Data sparseness: can't train word-specific classifiers

WSD Methods

- Supervised Machine Learning
- Thesaurus/Dictionary Methods
- Semi-Supervised Learning

Word Sense Disambiguation

Supervised Machine Learning

Supervised Machine Learning Approaches

- Supervised machine learning approach:
  - a training corpus of words tagged in context with their sense
  - used to train a classifier that can tag words in new text
- Summary of what we need:
  - the **tag set** (“sense inventory”)
  - the **training corpus**
  - A set of **features** extracted from the training corpus
  - A **classifier**

Supervised WSD 1: WSD Tags

- What's a tag?

  A dictionary sense?

- For example, for WordNet an instance of "bass" in a text has 8 possible tags or labels (bass1 through bass8).

8 senses of “bass” in WordNet

1. bass - (the lowest part of the musical range)
2. bass, bass part - (the lowest part in polyphonic music)
3. bass, basso - (an adult male singer with the lowest voice)
4. sea bass, bass - (flesh of lean-fleshed saltwater fish of the family Serranidae)
5. freshwater bass, bass - (any of various North American lean-fleshed freshwater fishes especially of the genus Micropterus)
6. bass, bass voice, basso - (the lowest adult male singing voice)
7. bass - (the member with the lowest range of a family of musical instruments)
8. bass - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Inventory of sense tags for *bass*

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…

Supervised WSD 2: Get a corpus

- Lexical sample task:
  - *Line-hard-serve* corpus - 4000 examples of each
  - *Interest* corpus - 2369 sense-tagged examples
- All words:
  - **Semantic concordance**: a corpus in which each open-class word is labeled with a sense from a specific dictionary/thesaurus.
    - SemCor: 234,000 words from Brown Corpus, manually tagged with WordNet senses
    - SENSEVAL-3 competition corpora - 2081 tagged word tokens

SemCor

<wf pos=PRP>**He**</wf>

<wf pos=VB lemma=recognize wnsn=4 lexsn=2:31:00::>**recognized**</wf>

<wf pos=DT>**the**</wf>

<wf pos=NN lemma=gesture wnsn=1 lexsn=1:04:00::>**gesture**</wf>

<punc>.</punc>

Supervised WSD 3: Extract feature vectors Intuition from Warren Weaver (1955):

"If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words…

But if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word…

The practical question is : ``What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?"

Feature vectors

- A simple representation for each observation
  (each instance of a target word)
  - **Vectors** of sets of feature/value pairs
  - Represented as a ordered list of values
  - These vectors represent, e.g., the window of words around the target

Two kinds of features in the vectors

- **Collocational** features and **bag-of-words** features
  - **Collocational**
    - Features about words at **specific** positions near target word
      - Often limited to just word identity and POS
  - **Bag-of-words**
    - Features about words that occur anywhere in the window (regardless of position)
      - Typically limited to frequency counts

Examples

- Example text (WSJ):

  An electric guitar and **bass** player stand off to one side not really part of the scene

- Assume a window of +/- 2 from the target

Examples

- Example text (WSJ)

  An electric $\boxed{\text{guitar}}\boxed{\text{and}}$ **bass** $\boxed{\text{player}}\boxed{\text{stand}}$ off to one side not really part of the scene,

- Assume a window of +/- 2 from the target

Collocational features

- Position-specific information about the words and collocations in window
- $\boxed{\text{guitar}}\boxed{\text{and}}$ bass $\boxed{\text{player}}\boxed{\text{stand}}$

$$[w_{i-2}, \mathrm{POS}_{i-2}, w_{i-1}, \mathrm{POS}_{i-1}, w_{i+1}, \mathrm{POS}_{i+1}, w_{i+2}, \mathrm{POS}_{i+2}, w_{i-2}^{i-1}, w_{i}^{i+1}]$$

```
[guitar, NN, and, CC, player, NN, stand, VB, and guitar, player stand]
```

- word 1,2,3 grams in window of ±3 is common

Bag-of-words features

- "an unordered set of words" – position ignored
- Counts of words occur within the window.
- First choose a vocabulary
- Then count how often each of those terms occurs in a given window
    - sometimes just a binary "indicator" 1 or 0

Co-Occurrence Example

- Assume we've settled on a possible vocabulary of 12 words in "bass" sentences:

[*fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band*]

- The vector for:

  guitar and bass player stand

  [0,0,0,1,0,0,0,0,0,0,1,0]

Word Sense Disambiguation

Classification

Classification: definition

- *Input*:
  - a word w and some features f
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
- *Output*: a predicted class $c \in C$

Classification Methods: Supervised Machine Learning

- *Input:*
  - a word w *in a text window* d *(which we'll call a "document")*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of m hand-labeled text windows again called "documents" $(d_1,c_1),\ldots,(d_m,c_m)$
- *Output:*
  - a learned classifier $\gamma: d \rightarrow c$

Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naive Bayes
  - Logistic regression
  - Neural Networks
  - Support-vector machines
  - k-Nearest Neighbors
  - …

Applying Naive Bayes to WSD

- P(c) is the prior probability of that sense
  - Counting in a labeled training set.
- P(w|c) conditional probability of a word given a particular sense
  - P(w|c) = count(w,c)/count(c)
- We get both of these from a tagged corpus like SemCor

- Can also generalize to look at other features besides words.
  - Then it would be P(f|c)
    - Conditional probability of a feature given a sense

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	fish smoked fish	f
	2	fish line	f
	3	fish haul smoked	f
	4	guitar jazz line	g
Test	5	line guitar jazz jazz	?

V = {fish, smoked, line, haul, guitar, jazz}

Priors:

$P(f) = \frac{3}{4}$

$P(g) = \frac{1}{4}$

Conditional Probabilities:

P(line|*f*) = (1+1) / (8+6) = 2/14

P(guitar|*f*) = (0+1) / (8+6) = 1/14

P(jazz|*f*) = (0+1) / (8+6) = 1/14

P(line|*g*) = (1+1) / (3+6) = 2/9

P(guitar|*g*) = (1+1) / (3+6) = 2/9

P(jazz|*g*) = (1+1) / (3+6) = 2/9

Choosing a class:

P(f|d5) $\propto$ 3/4 * 2/14 * (1/14)^2 * 1/14 ≈ 0.00003

P(g|d5) $\propto$ 1/4 * 2/9 * (2/9)^2 * 2/9 ≈ 0.0006

Word Sense Disambiguation

Evaluations and Baselines

WSD Evaluations and baselines

- Best evaluation: **extrinsic ('end-to-end', `task-based') evaluation**
  - Embed WSD algorithm in a task and see if you can do the task better!
- What we often do for convenience: **intrinsic evaluation**
  - Exact match **sense accuracy**
    - % of words tagged identically with the human-manual sense tags
  - Usually evaluate using **held-out data** from same labeled corpus
- Baselines
  - Most frequent sense
  - The Lesk algorithm

Most Frequent Sense

- WordNet senses are ordered in frequency order
- So "most frequent sense" in WordNet = "take the first sense"
- Sense frequencies come from the *SemCor* corpus

Freq	Synset	Gloss
338	plant^1, works, industrial plant	buildings for carrying on industrial labor
207	plant^2, flora, plant life	a living organism lacking the power of locomotion
2	plant^3	something planted secretly for discovery by another
0	plant^4	an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

Ceiling

- Human inter-annotator agreement
  - Compare annotations of two humans
  - On same data
  - Given same tagging guidelines
- Human agreements on all-words corpora with WordNet style senses
  - 75%-80%

Word Sense Disambiguation

Dictionary and Thesaurus Methods

The Simplified Lesk algorithm

- Let's disambiguate "**bank**" in this sentence:

  The **bank** can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

- given the following two WordNet senses:

$bank^1$	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the mortgage on my home"
$bank^2$	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

The Simplified Lesk algorithm

Choose sense with most word overlap between gloss and context
(not counting function words)

The **bank** can guarantee $\boxed{\text{deposits}}$ will eventually cover future tuition costs because it invests in adjustable-rate $\boxed{\text{mortgage}}$ securities.

bank^1	Gloss:	a financial institution that accepts $\boxed{\text{deposits}}$ and channels the money into lending activities
	Examples:	"he cashed a check at the bank", "that bank holds the $\boxed{\text{mortgage}}$ on my home"
bank^2	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	"they pulled the canoe up on the bank", "he sat on the bank of the river and watched the currents"

The Corpus Lesk algorithm

- Assumes we have some sense-labeled data (like SemCor)
- Take all the sentences with the relevant word sense:

  *These short, "streamlined" meetings usually are sponsored by local **banks**[1], Chambers of Commerce, trade associations, or other civic organizations.*

- Now add these to the gloss + examples for each sense, call it the "signature" of a sense.
- Choose sense with most word overlap between context and signature.

LESK'S ALGORITHM

Sense Bag: *contains the words in the definition of a candidate sense of the ambiguous word.*

Context Bag: *contains the words in the definition of each sense of each context word.*

E.g. "On burning ***coal*** we get ***ash***."

Ash

- Sense 1
  Trees of the olive family with pinnate leaves, thin furrowed bark and gray branches.
- Sense 2
  The ***solid*** residue left when ***combustible*** material is thoroughly ***burn***ed or oxidized.
- Sense 3
  To convert into ash

Coal

- Sense 1
  A piece of glowing carbon or ***burn***t wood.
- Sense 2
  charcoal.
- Sense 3
  A black ***solid combustible*** substance formed by the partial decomposition of vegetable matter without free access to air and under the influence of moisture and often increased pressure and temperature that is widely used as a fuel for ***burn***ing

In this case Sense 2 of ash would be the winner sense.

Corpus Lesk: IDF weighting

- Instead of just removing function words
  - Down-weights words that occur in every `document’ (gloss, example, etc)
  - These are generally function words, but is a more fine-grained measure
- Weigh each overlapping word by **inverse document frequency**

Corpus Lesk: IDF weighting

- Weigh each overlapping word by **inverse document frequency**
  - N is the total number of documents
  - df_i = "document frequency of word *i*"
  - = # of documents with word *I*

$$\mathrm{idf}_i = \log\left(\frac{N}{df_i}\right)$$

$$score(sense_i, context_j) = \sum_{w \in overlap(signature_i, context_j)} \mathrm{idf}_w$$

Walker's Algorithm

- A Thesaurus Based approach.
- ***Step 1:*** *For each sense of the target word find the thesaurus category to which that sense belongs.*
- ***Step 2:*** *Calculate the score for each sense by using the context words. A context words will add 1 to the score of the sense if the thesaurus category of the word matches that of the sense.*

  - *E.g. The money in this **bank** fetches an interest of 8% per annum*
  - Target word: ***bank***
  - Clue words from the context: ***money, interest, annum, fetch***

	Sense1: Finance	Sense2: Location
Money	+1	0
Interest	+1	0
Fetch	0	0
Annum	+1	0
Total	3	0

Context words add 1 to the sense when the topic of the word matches that of the sense

WSD Using Conceptual Density

- Select a sense based on the ***relatedness*** of that word-sense to the context.
- Relatedness is measured in terms of conceptual distance
  - (i.e. how close the concept represented by the ***word*** and the concept represented by its ***context words*** are)
- This approach uses a structured hierarchical semantic net (*WordNet*) for finding the conceptual distance.
- Smaller the conceptual distance higher will be the conceptual density.
  - (i.e. if all words in the context are strong indicators of a particular concept then that concept will have a higher density.)

Conceptual Density (Example)

- The dots in the figure represent the senses of the word to be disambiguated or the senses of the words in context.
- The CD formula will yield highest density for the sub-hierarchy containing more senses.
- The sense of W contained in the sub-hierarchy with the highest CD will be chosen.

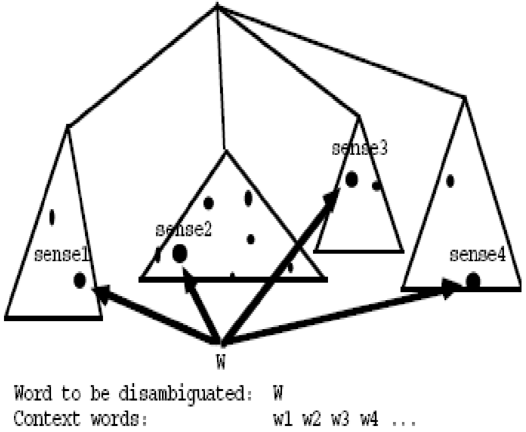
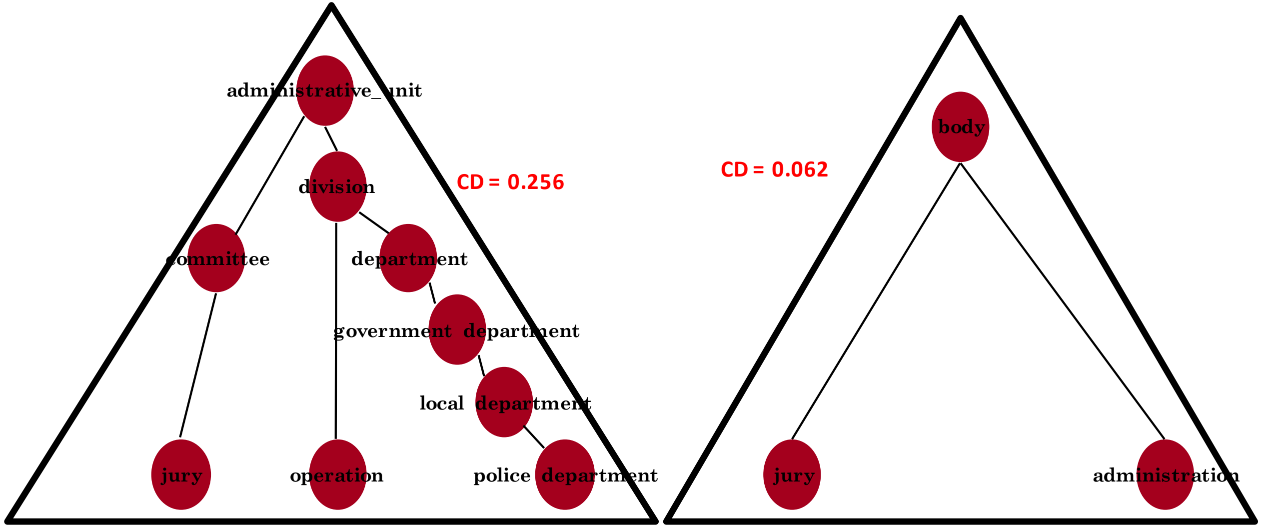


Figure 1: senses of a word in WordNet

Conceptual Density (Example)



The jury(2) praised the administration(3) and **operation** (8) of Atlanta Police Department(1)

Step 1: Make a lattice of the nouns in the context, their senses and hypernyms.

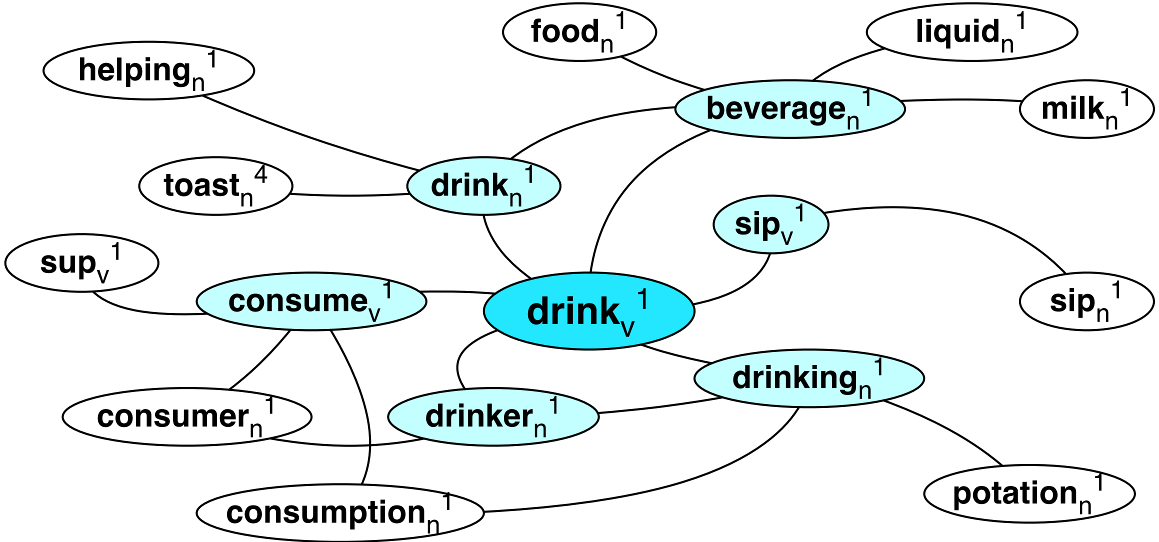
Step 2: Compute the conceptual density of resulting concepts (sub-hierarchies).

Step 3: The concept with highest CD is selected.

Step 4: Select the senses below the selected concept as the correct sense for the respective words.

Graph-based methods

- First, WordNet can be viewed as a graph
  - senses are nodes
  - relations (hypernymy, meronymy) are edges
  - Also add edge between word and unambiguous gloss words



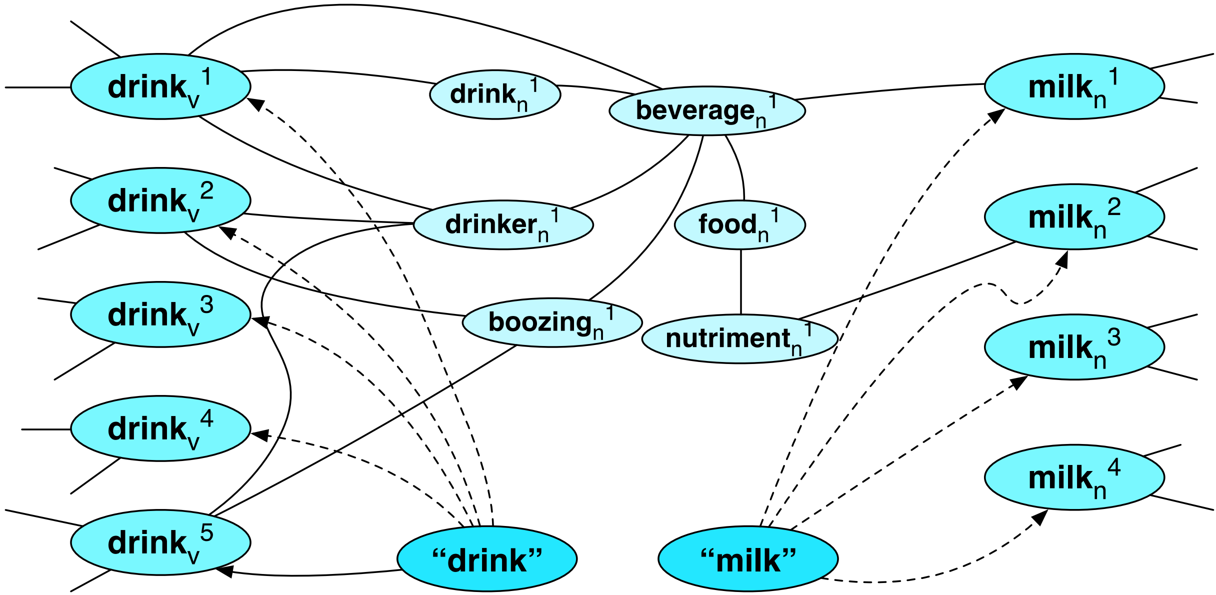
How to use the graph for WSD

- Insert target word and words in its sentential context into the graph, with directed edges to their senses

“She drank some milk”

- Now choose the *most central* sense

Add some probability to “drink” and “milk” and compute node with highest “pagerank”



WSD Using Random Walk Algorithm

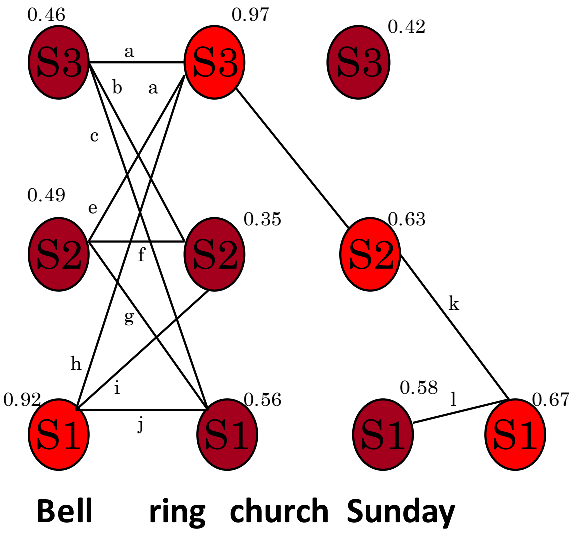
The **church bells** no longer **rung** on **Sundays**.

church
1: one of the groups of Christians who have their own beliefs and forms of worship
2: a place for public (especially Christian) worship
3: a service conducted in a church

bell
1: a hollow device made of metal that makes a ringing sound when struck
2: a push button at an outer door that gives a ringing or buzzing signal when pushed
3: the sound of a bell

ring
1: make a ringing sound
2: ring or echo with sound
3: make (bells) ring, often for the purposes of musical edification

Sunday
1: first day of the week; observed as a day of rest and worship by most Christians



Step 1: Add a vertex for each possible sense of each word in the text.

Step 2: Add weighted edges using definition based semantic similarity (Lesk's method).

Step 3: Apply graph based ranking algorithm to find score of each vertex (i.e. for each word sense).

Step 4: Select the vertex (sense) which has the highest score.

Word Sense Disambiguation

Semi-Supervised Learning

Semi-Supervised Learning

Problem: supervised and dictionary-based approaches require large hand-built resources

What if you don't have so much training data?

Solution: Bootstrapping

Generalize from a very small hand-labeled seed-set.

Bootstrapping

- For `bass`
  - Rely on “One sense per collocation” rule
    - A word reoccurring in collocation with the same word will almost surely have the same sense.

  - the word `play` occurs with the music sense of bass
  - the word `fish` occurs with the fish sense of bass

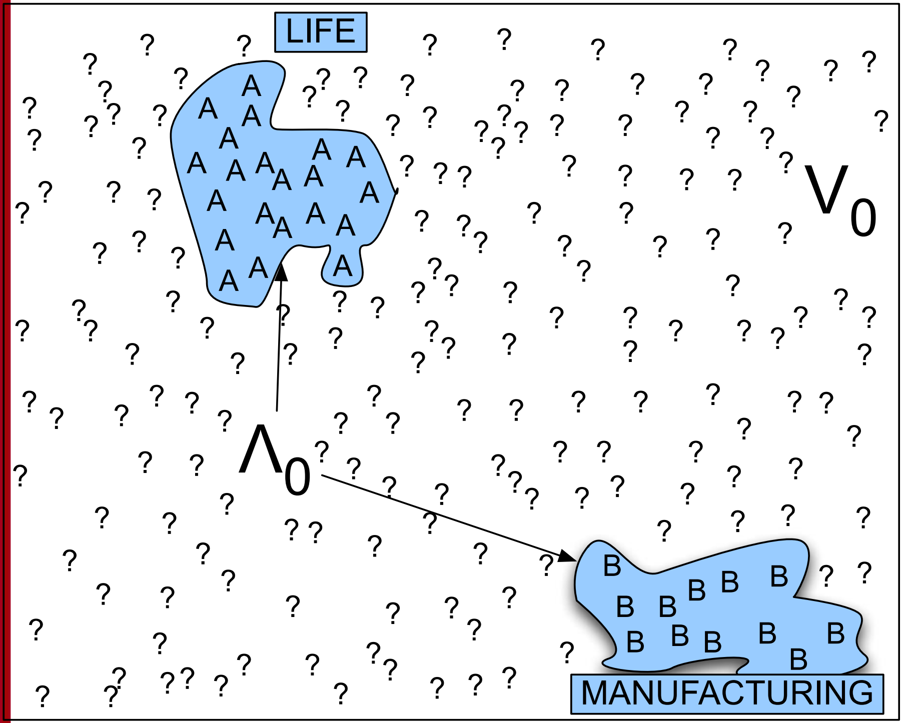
Sentences extracting using “fish” and “play”

We need more good teachers – right now, there are only a half a dozen who can **play** the free **bass** with ease. An electric guitar and **bass play**er stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.
The researchers said the worms spend part of their life cycle in such **fish** as Pacific salmon and striped **bass** and Pacific rockfish or snapper. And it all started when **fish**ermen decided the striped **bass** in Lake Mead were too skinny.

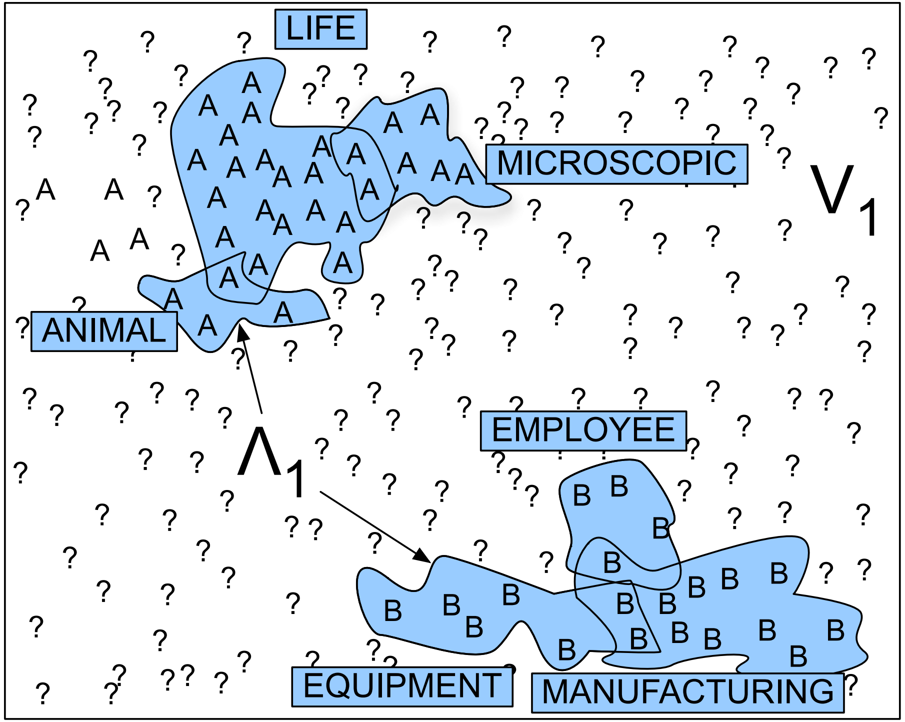
Summary: generating seeds

1) Hand labeling
2) “One sense per collocation”:
   - A word reoccurring in collocation with the same word will almost surely have the same sense.
3) “One sense per discourse”:
   - The sense of a word is highly consistent within a document - Yarowsky (1995)
   - (At least for non-function words, and especially topic-specific words)

Stages in the Yarowsky bootstrapping algorithm for the word “plant”



(a)



(b)

Summary

- Word Sense Disambiguation: choosing correct sense in context
- Applications: MT, QA, etc.
- Three classes of Methods
  - Supervised Machine Learning: Naive Bayes classifier
  - Thesaurus/Dictionary Methods
  - Semi-Supervised Learning
- Main intuition
  - There is lots of information in a word's context
  - Simple algorithms based just on word counts can be surprisingly good