Word Sense Disambiguation - II

Pawan Goyal

CSE, IIT Kharagpur

Week 8, Lecture 4

Minimally Supervised WSD - Yarowsky

- Annotations are expensive!
- "Bootstrapping" or co-training
  - Start with (small) seed, learn decision list
  - Use decision list to label rest of corpus
  - Retain 'confident' labels, treat as annotated data to learn new decision list
  - Repeat ...

Minimally Supervised WSD - Yarowsky

- Annotations are expensive!
- “Bootstrapping” or co-training
  - Start with (small) seed, learn decision list
  - Use decision list to label rest of corpus
  - Retain ‘confident’ labels, treat as annotated data to learn new decision list
  - Repeat ...
- Heuristics (derived from observation):
  - One sense per discourse
  - One sense per collocation

More about heuristics

One Sense per Discourse

- A word tends to preserve its meaning across all its occurrences in a given discourse

More about heuristics

One Sense per Discourse

- A word tends to preserve its meaning across all its occurrences in a given discourse

One Sense per Collocation

- A word tends to preserve its meaning when used in the same collocation
  - Strong for adjacent collocations
  - Weaker as the distance between the words increases

Yarowsky's Method

Example

- Disambiguating plant (industrial sense) vs. plant (living thing sense)
- Think of seed features for each sense
  - Industrial sense: co-occurring with 'manufacturing'
  - Living thing sense: co-occurring with 'life'
- Use 'one sense per collocation' to build initial decision list classifier
- Treat results (having high probability) as annotated data, train new decision list classifier, iterate

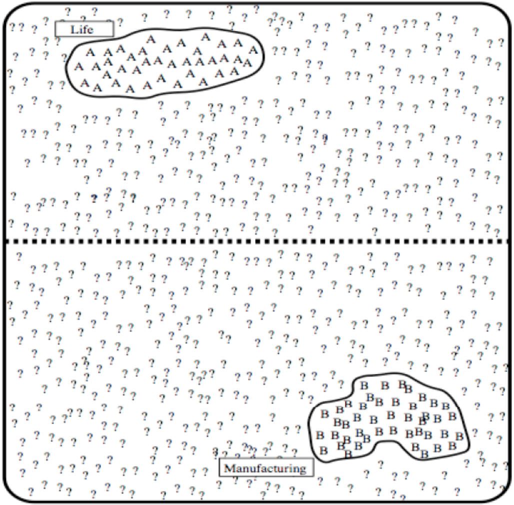
Yarowsky's Method: Example

used to strain microscopic plant life from the
zonal distribution of plant life .
close-up studies of plant life and natural
too rapid growth of aquatic plant life in water
the proliferation of plant and animal life
establishment phase of the plant virus life cycle
that divide life into plant and animal kingdom
many dangers to plant and animal life
mammals . Animal and plant life are delicately

automated manufacturing plant in Fremont
vast manufacturing plant and distribution
chemical manufacturing plant , producing viscose
keep a manufacturing plant profitable without
computer manufacturing plant and adjacent
discovered at a St. Louis plant manufacturing
copper manufacturing plant found that they
copper wire manufacturing plant , for example
s cement manufacturing plant in Alpena

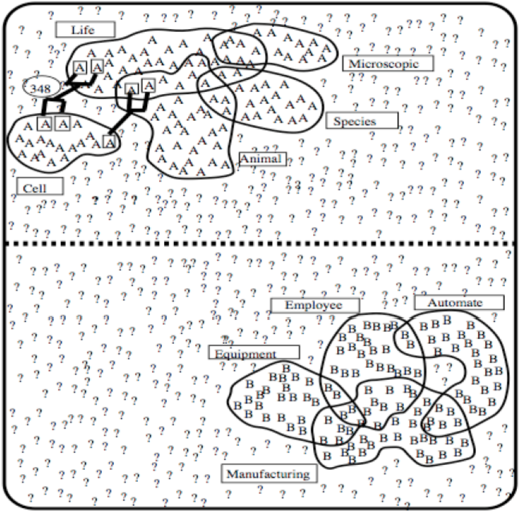
vinyl chloride monomer plant , which is
molecules found in plant and animal tissue
Nissan car and truck plant in Japan is
and Golgi apparatus of plant and animal cells
union responses to plant closures .
cell types found in the plant kingdom are
company said the plant is still operating
Although thousands of plant and animal species
animal rather than plant tissues can be

Yarowsky's Method: Example



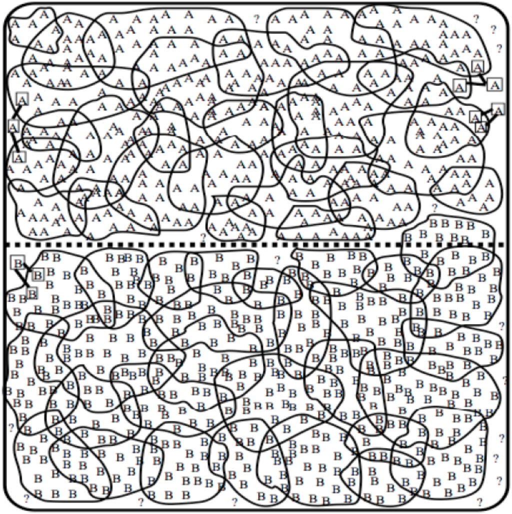
Initial state after use of seed rules

Yarowsky's Method: Example



Intermediate state

Yarowsky's Method: Example



Final state

Yarowsky's Method

Termination

- Stop when
  - Error on training data is less than a threshold
  - No more training data is covered
- Use final decision list for WSD

Advantages

- Accuracy is about as good as a supervised algorithm
- Bootstrapping: far less manual effort

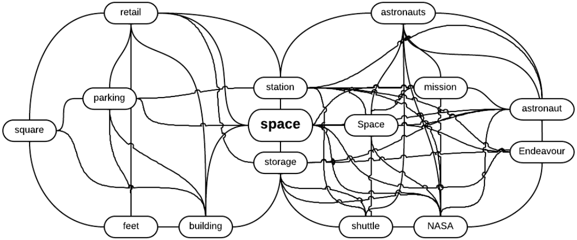
HyperLex

Key Idea: Word Sense Induction

- Instead of using “dictionary defined senses”, extract the “senses from the corpus” itself
- These “corpus senses” or “uses” correspond to clusters of similar contexts for a word.

Key Idea: Word Sense Induction

- Instead of using "dictionary defined senses", extract the "senses from the corpus" itself
- These "corpus senses" or "uses" correspond to clusters of similar contexts for a word.



Detecting Root Hubs

- Different uses of a target word form highly interconnected bundles (or high density components)
- In each high density component one of the nodes (hub) has a higher degree than the others.
- **Step 1:** Construct co-occurrence graph, G.

HyperLex

Detecting Root Hubs

- Different uses of a target word form highly interconnected bundles (or high density components)
- In each high density component one of the nodes (hub) has a higher degree than the others.
- **Step 1:** Construct co-occurrence graph, G.
- **Step 2:** Arrange nodes in G in decreasing order of degree.

HyperLex

Detecting Root Hubs

- Different uses of a target word form highly interconnected bundles (or high density components)
- In each high density component one of the nodes (hub) has a higher degree than the others.
- **Step 1:** Construct co-occurrence graph, G.
- **Step 2:** Arrange nodes in G in decreasing order of degree.
- **Step 3:** Select the node from G which has the highest degree. This node will be the hub of the first high density component.

HyperLex

Detecting Root Hubs

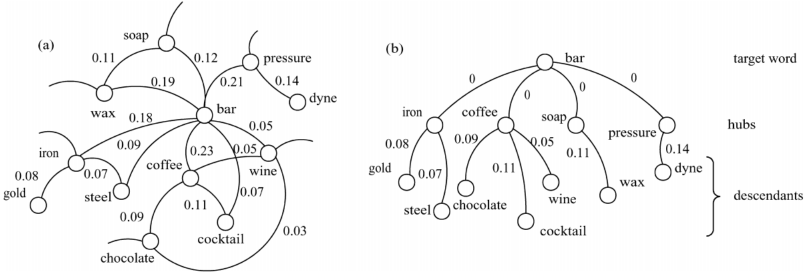
- Different uses of a target word form highly interconnected bundles (or high density components)
- In each high density component one of the nodes (hub) has a higher degree than the others.
- **Step 1:** Construct co-occurrence graph, G.
- **Step 2:** Arrange nodes in G in decreasing order of degree.
- **Step 3:** Select the node from G which has the highest degree. This node will be the hub of the first high density component.
- **Step 4:** Delete this hub and all its neighbors from G.

HyperLex

Detecting Root Hubs

- Different uses of a target word form highly interconnected bundles (or high density components)
- In each high density component one of the nodes (hub) has a higher degree than the others.
- **Step 1:** Construct co-occurrence graph, G.
- **Step 2:** Arrange nodes in G in decreasing order of degree.
- **Step 3:** Select the node from G which has the highest degree. This node will be the hub of the first high density component.
- **Step 4:** Delete this hub and all its neighbors from G.
- **Step 5:** Repeat Step 3 and 4 to detect the hubs of other high density components

HyperLex: Detecting Root Hubs



Delineating Components

- Attach each node to the root hub closest to it.
- The distance between two nodes is measured as the smallest sum of weights of the edges on the paths linking them.

Delineating Components

- Attach each node to the root hub closest to it.
- The distance between two nodes is measured as the smallest sum of weights of the edges on the paths linking them.

Computing distance between two nodes w_i and w_j

$$w_{ij} = 1 - max\{P(w_i|w_j), P(w_j|w_i)\}$$

where $P(w_i|w_j) = \frac{freq_{ij}}{freq_j}$

Disambiguation

- Let $W = (w_1, w_2, \ldots, w_i, \ldots, w_n)$ be a context in which w_i is an instance of our target word.
- Let w_i has k hubs in its minimum spanning tree

Disambiguation

- Let $W = (w_1, w_2, \ldots, w_i, \ldots, w_n)$ be a context in which w_i is an instance of our target word.
- Let w_i has k hubs in its minimum spanning tree
- A score vector s is associated with each $w_j \in W (j \neq i)$, such that s_k represents the contribution of the kth hub as:

$$\begin{aligned} s_k &= \frac{1}{1 + d(h_k, w_j)} \text{if } h_k \text{ is an ancestor of } w_j \\ s_i &= 0 \text{ otherwise.} \end{aligned}$$

- All score vectors associated with all $w_j \in W (j \neq i)$ are summed up
- The hub which receives the maximum score is chosen as the most appropriate sense