Word Sense Disambiguation

Ling571
Deep Processing Techniques for NLP
March 3, 2014

Distributional Similarity Questions

- What is the right neighborhood?
  - What is the context?
- How should we weight the features?
- How can we compute similarity between vectors?

Feature Vector Design

- Window size:
  - How many words in the neighborhood?
    - Tradeoff:
      - +/- 500 words: 'topical context'
      - +/- 1 or 2 words: collocations, predicate-argument
      - Only words in some grammatical relation
        - Parse text (dependency)
        - Include subj-verb; verb-obj; adj-mod
          - NxR vector: word x relation

Example Lin Relation Vector

...	*pobj-of*, inside	*pobj-of*, into	...	*nmod-of*, abnormality	*nmod-of*, anemia	*nmod-of*, architecture	...	*obj-of*, attack	*obj-of*, call	*obj-of*, come from	*obj-of*, decorate
	16	30		3	8	1		6	11	3	2

Weighting Features

- Baseline: Binary (0/1)
  - Minimally informative
  - Can't capture intuition that frequent features informative
- Frequency or Probability:

$$P(f \mid w) = \frac{count(f, w)}{count(w)}$$

  - Better but,
  - Can overweight a priori frequent features
    - Chance cooccurrence

Pointwise Mutual Information

$$assoc_{PMI}(w,f) = \log_2 \frac{P(w,f)}{P(w)P(f)}$$

PMI:

- Contrasts observed cooccurrence
- With that expected by chance (if independent)
- Generally only use positive values
- Negatives inaccurate unless corpus huge

Vector Similarity

- Euclidean or Manhattan distances:
  - Too sensitive to extreme values

- Dot product: $sim_{dot-product}(\vec{v},\vec{w}) = \vec{v} \bullet \vec{w} = \sum_{i=1}^{N} v_i \times w_i$
  - Favors long vectors:
    - More features or higher values

- Cosine: $sim_{\cos ine}(\vec{v},\vec{w}) = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$

Distributional Similarity for Word Sense Disambiguation

Schutze's Word Space

- Build a co-occurrence matrix
  - Restrict Vocabulary to 4 letter sequences
    - Similar effect to stemming
    - Exclude Very Frequent - Articles, Affixes
  - Entries in 5000-5000 Matrix
    - Apply Singular Value Decomposition (SVD)
    - Reduce to 97 dimensions
- Word Context
  - 4grams within 1001 Characters
  - Sum & Normalize Vectors for each 4gram
  - Distances between Vectors by dot product

Schutze's Word Space

- Word Sense Disambiguation
  - Context Vectors of All Instances of Word
  - Automatically Cluster Context Vectors
  - Hand-label Clusters with Sense Tag
  - Tag New Instance with Nearest Cluster

There are more kinds of plants and animals in the rainforests than anywhere else on Earth. Over half of the millions of known species of plants and animals live in the rainforest. Many are found nowhere else. There are even plants and animals in the rainforest that we have not yet discovered.

Biological Example

The Paulus company was founded in 1938. Since those days the product range has been the subject of constant expansions and is brought up continuously to correspond with the state of the art. We’re engineering, manufacturing and commissioning world-wide ready-to-run plants packed with our comprehensive know-how. Our Product Range includes pneumatic conveying systems for carbon, carbide, sand, lime andmany others. We use reagent injection in molten metal for the…

Industrial Example

Label the First Use of “Plant”

Sense Selection in "Word Space"

- Build a Context Vector
  - 1,001 character window - Whole Article
- Compare Vector Distances to Sense Clusters
  - Only 3 Content Words in Common
  - Distant Context Vectors
  - Clusters - Build Automatically, Label Manually
- Result: 2 Different, Correct Senses
  - 92% on Pair-wise tasks

Odd Cluster Examples

- The “Ste.” Cluster:
  - Dry Oyster Whisky Hot Float Ice

Odd Cluster Examples

- The “Ste.” Cluster:
  - Dry Oyster Whisky Hot Float Ice
  - Why? – River name

Odd Cluster Examples

- The “Ste.” Cluster:
  - Dry Oyster Whisky Hot Float Ice
  - Why? – River name
    - Learning the Corpus, not the Sense
- Keeping cluster:
  - Bring Hoping Wiping Could Should Some Them Rest

Taxonomy of Contextual Information

- Topical Content
- Word Associations
- Syntactic Constraints
- Selectional Preferences
- World Knowledge & Inference

The Question of Context

- Shared Intuition:
  - Context -> Sense
- Area of Disagreement:
  - What is context?
- Wide vs Narrow Window
- Word Co-occurrences
- Best model, best weighting
  - Still active focus of research

Minimally Supervised WSD

- Yarowsky's algorithm (1995)

- Bootstrapping approach:
  - Use small labeled seedset to iteratively train

- Builds on 2 key insights:
  - One Sense Per Discourse
    - Word appearing multiple times in text has same sense
    - Corpus of 37232 bass instances: always single sense

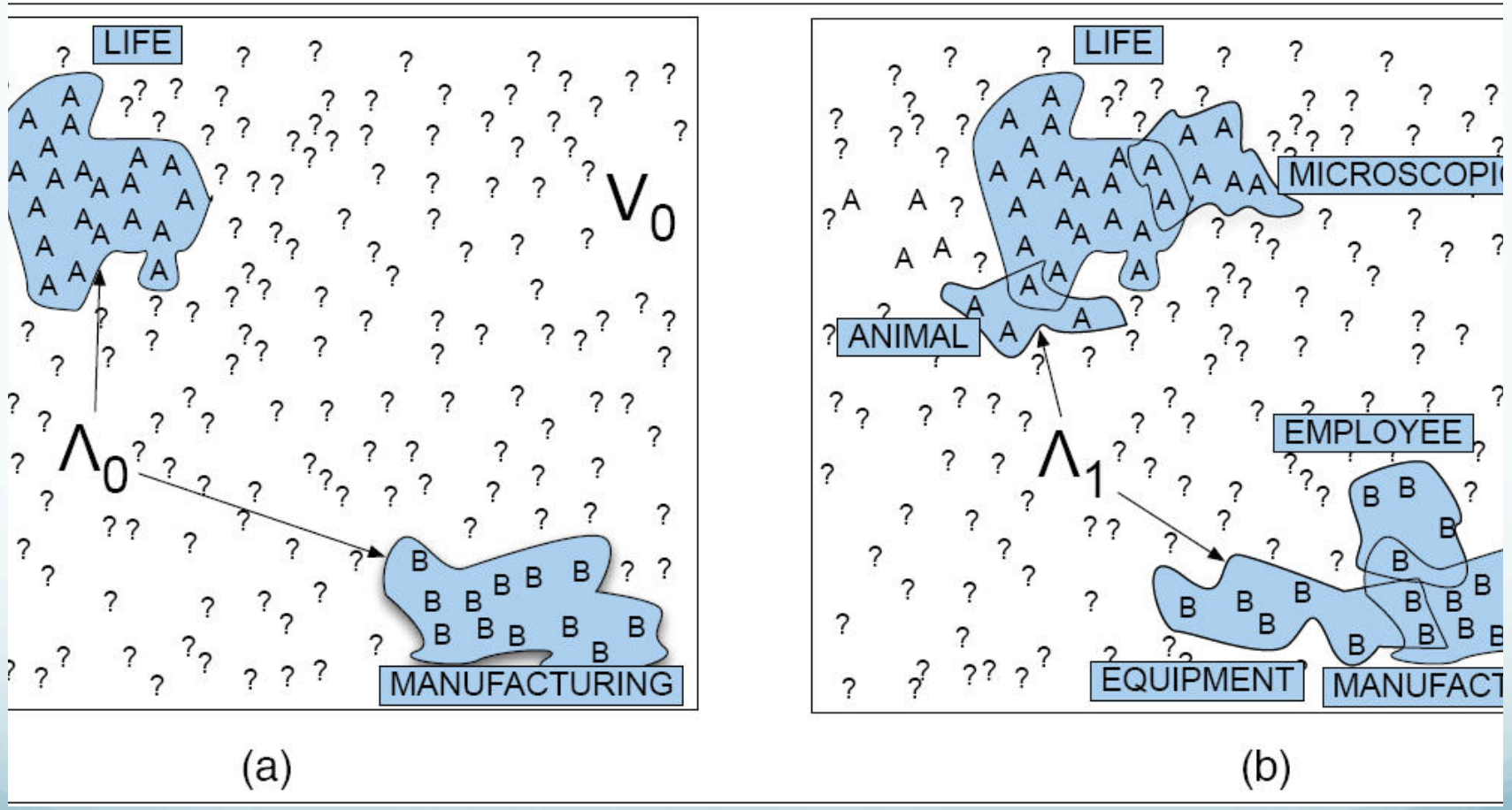
  - One Sense Per Collocation
    - Local phrases select single sense
      - Fish -> Bass[1]
      - Play -> Bass[2]

Yarowsky's Algorithm

- Training Decision Lists
  - 1. Pick Seed Instances & Tag
  - 2. Find Collocations: Word Left, Word Right, Word ±K
    - (A) Calculate Informativeness on Tagged Set,
      - Order: $$abs(\log \frac{P(Sense_1 \mid Collocation)}{P(Sense_2 \mid Collocation)})$$
    - (B) Tag New Instances with Rules
    - (C) Apply 1 Sense/Discourse
    - (D) If Still Unlabeled, Go To 2
  - 3. Apply 1 Sense/Discourse
- Disambiguation: First Rule Matched

Initial decision list for *plant* (abbreviated)		
LogL	Collocation	Sense
8.10	***plant* life**	⇒ A
7.58	**manufacturing *plant***	⇒ B
7.39	**life (within ±2-10 words)**	⇒ A
7.20	**manufacturing (in ±2-10 words)**	⇒ B
6.27	**animal (within ±2-10 words)**	⇒ A
4.70	**equipment (within ±2-10 words)**	⇒ B
4.39	**employee (within ±2-10 words)**	⇒ B
4.30	**assembly *plant***	⇒ B
4.10	***plant* closure**	⇒ B
3.52	***plant* species**	⇒ A
3.48	**automate (within ±2-10 words)**	⇒ B
3.45	**microscopic *plant***	⇒ A
	...	

Iterative Updating



There are more kinds of plants and animals in the rainforests than anywhere else on Earth. Over half of the millions of known species of plants and animals live in the rainforest. Many are found nowhere else. There are even plants and animals in the rainforest that we have not yet discovered.

Biological Example

The Paulus company was founded in 1938. Since those days the product range has been the subject of constant expansions and is brought up continuously to correspond with the state of the art. We’re engineering, manufacturing and commissioning world-wide ready-to-run plants packed with our comprehensive know-how. Our Product Range includes pneumatic conveying systems for carbon, carbide, sand, lime andmany others. We use reagent injection in molten metal for the…

Industrial Example

Label the First Use of “Plant”

Sense Choice With Collocational Decision Lists

- Create Initial Decision List
  - Rules Ordered by $abs(\log \frac{P(Sense_1 \mid Collocation)}{P(Sense_2 \mid Collocation)})$
- Check nearby Word Groups (Collocations)
  - Biology: "Animal" in ± 2-10 words
  - Industry: "Manufacturing" in ± 2-10 words
- Result: Correct Selection
  - 95% on Pair-wise tasks

Naïve Bayes' Approach

- Supervised learning approach
  - Input: feature vector X label
- Best sense = most probable sense given f

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

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

Naïve Bayes' Approach

- Issue:
  - Data sparseness: full feature vector rarely seen
- "Naïve" assumption:
  - Features independent given sense

$$P(\vec{f} \mid s) \approx \prod_{j=1}^{n} P(f_j \mid s)$$

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

Training NB Classifier

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

- Estimate P(s):
  - Prior

$$P(s_i) = \frac{count(s_i, w_j)}{count(w_j)}$$

- Estimate $P(f_j|s)$ $\quad P(f_j \mid s) = \frac{count(f_j, s)}{count(s)}$
- Issues:
  - Underflow => log prob
  - Sparseness => smoothing