Word Sense Disambiguation

Ling571
Deep Processing Techniques for NLP
February 28, 2011

Word Sense Disambiguation

- Robust Approaches
  - Similarity-based approaches
    - Thesaurus-based techniques
      - Resnik/Lin similarity
    - Unsupervised, distributional approaches
      - Word-space (Infomap)
  - Why they work
  - Why they don't

Resnik's Similarity Measure

- Information content of node:
  - $IC(c) = -\log P(c)$
- Least common subsumer (LCS):
  - Lowest node in hierarchy subsuming 2 nodes
- Similarity measure:
  - $sim_{RESNIK}(c_1,c_2) = -\log P(LCS(c_1,c_2))$
- Issue:
  - Not content, but difference between node & LCS

$$sim_{Lin}(c_1,c_2) = \frac{2 \times \log P(LCS(c_1,c_2))}{\log P(c_1) + \log P(c_2)}$$

Application to WSD

- Calculate Informativeness
  - For Each Node in WordNet:

Application to WSD

- Calculate Informativeness
  - For Each Node in WordNet:
    - Sum occurrences of concept and all children
    - Compute IC

Application to WSD

- Calculate Informativeness
  - For Each Node in WordNet:
    - Sum occurrences of concept and all children
    - Compute IC
- Disambiguate with WordNet

Application to WSD

- Calculate Informativeness
  - For Each Node in WordNet:
    - Sum occurrences of concept and all children
    - Compute IC
- Disambiguate with WordNet
  - Assume set of words in context
    - E.g. {plants, animals, rainforest, species} from article

Application to WSD

- Calculate Informativeness
  - For Each Node in WordNet:
    - Sum occurrences of concept and all children
    - Compute IC
- Disambiguate with WordNet
  - Assume set of words in context
    - E.g. {plants, animals, rainforest, species} from article
  - Find Most Informative Subsumer for each pair, I
    - Find LCS for each pair of senses, pick highest similarity

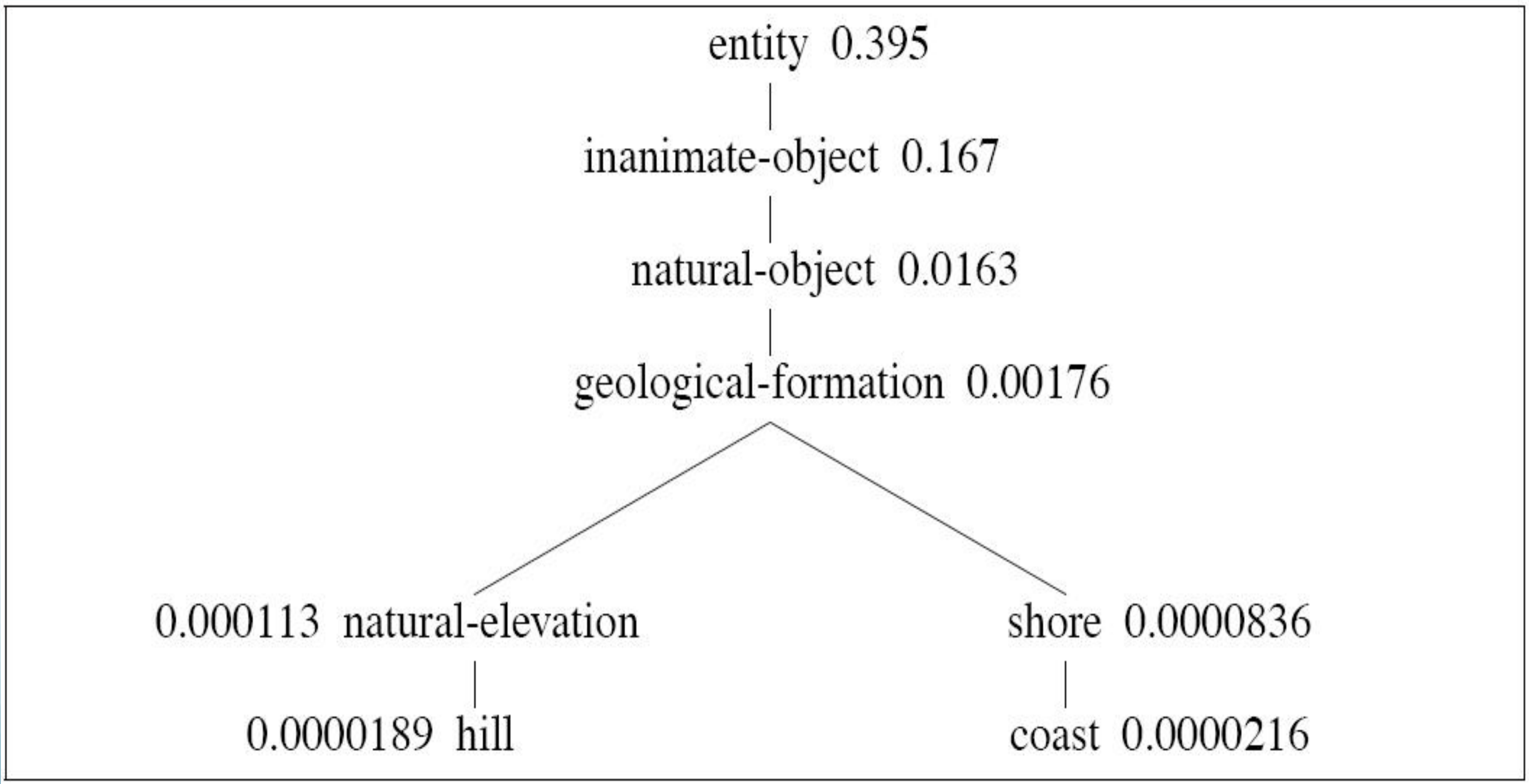
Application to WSD

- Calculate Informativeness
  - For Each Node in WordNet:
    - Sum occurrences of concept and all children
    - Compute IC
- Disambiguate with WordNet
  - Assume set of words in context
    - E.g. {plants, animals, rainforest, species} from article
  - Find Most Informative Subsumer for each pair, I
    - Find LCS for each pair of senses, pick highest similarity
  - For each subsumed sense, Vote += I

Application to WSD

- Calculate Informativeness
  - For Each Node in WordNet:
    - Sum occurrences of concept and all children
    - Compute IC
- Disambiguate with WordNet
  - Assume set of words in context
    - E.g. {plants, animals, rainforest, species} from article
  - Find Most Informative Subsumer for each pair, I
    - Find LCS for each pair of senses, pick highest similarity
  - For each subsumed sense, Vote += I
  - Select Sense with Highest Vote

IC Example



entity 0.395
inanimate-object 0.167
natural-object 0.0163
geological-formation 0.00176
0.000113 natural-elevation
shore 0.0000836
0.0000189 hill
coast 0.0000216

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 Labeling Under WordNet

- Use Local Content Words as Clusters
  - Biology: Plants, Animals, Rainforests, species…
  - Industry: Company, Products, Range, Systems…

Sense Labeling Under WordNet

- Use Local Content Words as Clusters
  - Biology: Plants, Animals, Rainforests, species…
  - Industry: Company, Products, Range, Systems…
- Find Common Ancestors in WordNet

Sense Labeling Under WordNet

- Use Local Content Words as Clusters
  - Biology: Plants, Animals, Rainforests, species…
  - Industry: Company, Products, Range, Systems…
- Find Common Ancestors in WordNet
  - Biology: Plants & Animals isa Living Thing

Sense Labeling Under WordNet

- Use Local Content Words as Clusters
  - Biology: Plants, Animals, Rainforests, species…
  - Industry: Company, Products, Range, Systems…
- Find Common Ancestors in WordNet
  - Biology: Plants & Animals isa Living Thing
  - Industry: Product & Plant isa Artifact isa Entity

Sense Labeling Under WordNet

- Use Local Content Words as Clusters
  - Biology: Plants, Animals, Rainforests, species…
  - Industry: Company, Products, Range, Systems…
- Find Common Ancestors in WordNet
  - Biology: Plants & Animals isa Living Thing
  - Industry: Product & Plant isa Artifact isa Entity
  - Use Most Informative
- Result: Correct Selection

Thesaurus Similarity Issues

Thesaurus Similarity Issues

- Coverage:
  - Few languages have large thesauri

Thesaurus Similarity Issues

- Coverage:
  - Few languages have large thesauri
  - Few languages have large sense tagged corpora

Thesaurus Similarity Issues

- Coverage:
  - Few languages have large thesauri
  - Few languages have large sense tagged corpora

- Thesaurus design:
  - Works well for noun IS-A hierarchy

Thesaurus Similarity Issues

- Coverage:
  - Few languages have large thesauri
  - Few languages have large sense tagged corpora

- Thesaurus design:
  - Works well for noun IS-A hierarchy
  - Verb hierarchy shallow, bushy, less informative

Distributional Similarity

- Unsupervised approach:
  - Clustering, WSD, automatic thesaurus enrichment

Distributional Similarity

- Unsupervised approach:
  - Clustering, WSD, automatic thesaurus enrichment
- Insight:
  - “You shall know a word by the company it keeps!”
    - (Firth, 1957)

Distributional Similarity

- Unsupervised approach:
  - Clustering, WSD, automatic thesaurus enrichment
- Insight:
  - "You shall know a word by the company it keeps!"
    - (Firth, 1957)
  - A bottle of *tezguino* is on the table.
  - Everybody likes *tezguino*.
  - *Tezguino* makes you drunk.
  - We make *tezguino* from corn.

Distributional Similarity

- Unsupervised approach:
  - Clustering, WSD, automatic thesaurus enrichment
- Insight:
  - "You shall know a word by the company it keeps!"
    - (Firth, 1957)
  - A bottle of *tezguino* is on the table.
  - Everybody likes *tezguino*.
  - *Tezguino* makes you drunk.
  - We make *tezguino* from corn.
- Tezguino: corn-based, alcoholic beverage

Distributional Similarity

- Represent ‘company’ of word such that similar words will have similar representations
  - ‘Company’ = context

Distributional Similarity

- Represent ‘company’ of word such that similar words will have similar representations
  - ‘Company’ = context
- Word represented by context feature vector
  - Many alternatives for vector

Distributional Similarity

- Represent ‘company’ of word such that similar words will have similar representations
  - ‘Company’ = context
- Word represented by context feature vector
  - Many alternatives for vector
- Initial representation:
  - ‘Bag of words’ binary feature vector

Distributional Similarity

- Represent 'company' of word such that similar words will have similar representations
  - 'Company' = context
- Word represented by context feature vector
  - Many alternatives for vector
- Initial representation:
  - 'Bag of words' binary feature vector
  - Feature vector length N, where N is of vocabulary
    - f_i= 1 if $word_i$ within window of *w*, 0 o.w.

Binary Feature Vector

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

Distributional Similarity Questions

Distributional Similarity Questions

- What is the right neighborhood?
  - What is the context?

Distributional Similarity Questions

- What is the right neighborhood?
  - What is the context?

- How should we weight the features?

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:

Feature Vector Design

- Window size:
  - How many words in the neighborhood?
    - Tradeoff:
      - +/- 500 words

Feature Vector Design

- Window size:
  - How many words in the neighborhood?
    - Tradeoff:
      - +/- 500 words: ‘topical context’

Feature Vector Design

- Window size:
  - How many words in the neighborhood?
    - Tradeoff:
      - +/- 500 words: ‘topical context’
      - +/- 1 or 2 words:

Feature Vector Design

- Window size:
  - How many words in the neighborhood?
    - Tradeoff:
      - +/- 500 words: ‘topical context’
      - +/- 1 or 2 words: collocations, predicate-argument

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*, decora
	16	30		3	8	1		6	11	3	2

Weighting Features

- Baseline: Binary (0/1)

Weighting Features

- Baseline: Binary (0/1)
  - Minimally informative
  - Can’t capture intuition that frequent features informative

Weighting Features

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

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

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,

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)

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:

Vector Similarity

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

Vector Similarity

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

- Dot product: $sim_{dot-product}(\vec{v},\vec{w}) = \vec{v} \bullet \vec{w} = \sum_{i=1}^{N} v_i \times w_i$

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

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: $sim_{cosine}(\vec{v},\vec{w}) = \dfrac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$

Schutze's Word Space

- Build a co-occurrence matrix

Schutze's Word Space

- Build a co-occurrence matrix
  - Restrict Vocabulary to 4 letter sequences

Schutze's Word Space

- Build a co-occurrence matrix
  - Restrict Vocabulary to 4 letter sequences
    - Similar effect to stemming
    - Exclude Very Frequent - Articles, Affixes

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

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

Schutze's Word Space

- Word Sense Disambiguation
  - Context Vectors of All Instances of Word

Schutze's Word Space

- Word Sense Disambiguation
  - Context Vectors of All Instances of Word
  - Automatically Cluster Context Vectors

Schutze's Word Space

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

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

Sense Selection in "Word Space"

- Build a Context Vector
  - 1,001 character window - Whole Article

Sense Selection in "Word Space"

- Build a Context Vector
  - 1,001 character window - Whole Article
- Compare Vector Distances to Sense Clusters

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

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

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
    - Uninformative: Wide context misses verb sense

The Question of Context

- Shared Intuition:
  - Context -> Sense
- Area of Disagreement:
  - What is context?
- Wide vs Narrow Window
- Word Co-occurrences

Taxonomy of Contextual Information

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

Limits of Wide Context

- Comparison of Wide-Context Techniques (LTV ‘93)
  - Neural Net, Context Vector, Bayesian Classifier, Simulated Annealing

Limits of Wide Context

- Comparison of Wide-Context Techniques (LTV ‘93)
  - Neural Net, Context Vector, Bayesian Classifier, Simulated Annealing
    - Results: 2 Senses - 90+%; 3+ senses ~ 70%

Limits of Wide Context

- Comparison of Wide-Context Techniques (LTV '93)
  - Neural Net, Context Vector, Bayesian Classifier, Simulated Annealing
    - Results: 2 Senses - 90+%; 3+ senses ~ 70%
      - Nouns: 92%; Verbs: 69%

Limits of Wide Context

- Comparison of Wide-Context Techniques (LTV '93)
  - Neural Net, Context Vector, Bayesian Classifier, Simulated Annealing
    - Results: 2 Senses - 90+%; 3+ senses ~ 70%
      - Nouns: 92%; Verbs: 69%
    - People: Sentences ~100%; Bag of Words: ~70%

Limits of Wide Context

- Comparison of Wide-Context Techniques (LTV ‘93)
  - Neural Net, Context Vector, Bayesian Classifier, Simulated Annealing
    - Results: 2 Senses - 90+%; 3+ senses ~ 70%
      - Nouns: 92%; Verbs: 69%
    - People: Sentences ~100%; Bag of Words: ~70%
- Inadequate Context

Limits of Wide Context

- Comparison of Wide-Context Techniques (LTV '93)
  - Neural Net, Context Vector, Bayesian Classifier, Simulated Annealing
    - Results: 2 Senses - 90+%; 3+ senses ~ 70%
      - Nouns: 92%; Verbs: 69%
    - People: Sentences ~100%; Bag of Words: ~70%
- Inadequate Context
- Need Narrow Context
  - Local Constraints Override
  - Retain Order, Adjacency

Interactions Below the Surface

- Constraints Not All Created Equal
  - “The Astronomer Married the Star”

Interactions Below the Surface

- Constraints Not All Created Equal
  - “The Astronomer Married the Star”
  - Selectional Restrictions Override Topic

Interactions Below the Surface

- Constraints Not All Created Equal
  - “The Astronomer Married the Star”
  - Selectional Restrictions Override Topic
- No Surface Regularities
  - “The emigration/immigration bill guaranteed passports to all Soviet citizens

Interactions Below the Surface

- Constraints Not All Created Equal
  - “The Astronomer Married the Star”
  - Selectional Restrictions Override Topic
- No Surface Regularities
  - “The emigration/immigration bill guaranteed passports to all Soviet citizens
  - No Substitute for Understanding