Lexical Semantics & WSD

Ling571
Deep Processing Techniques for NLP
February 24, 2016

Roadmap

- Distributional models
  - Compression
  - Integration
- Dictionary-based models
- Thesaurus-based similarity models
  - WordNet
  - Distance & Similarity in a Thesaurus
- Classifier models

Curse of Dimensionality

- Vector representations:
  - Sparse
  - Very high dimensional:
    - # words in vocabulary
    - # relations x # words, etc
- Google1T5 corpus:
  - 1M x 1M matrix: < 0.05% non-zero values
- Computationally hard to manage
  - Lots of zeroes
  - Can miss underlying relations

Reducing Dimensionality

- Feature selection:
  - Desirable traits:
    - High frequency
    - High variance
- Filtering:
  - Can exclude terms with too few occurrences
  - Can include only top X most frequent terms
  - Chi-squared selection
- Cautions:
  - Feature correlations
  - Joint feature selection complex, expensive

Reducing Dimensionality

- Projection into lower dimensional space:
  - Principal Components Analysis (PCA), Locality Preserving Projections (LPP), Singular Value Decomposition, etc
- Create new lower dimensional space that
  - Preserves distances between data points
    - Keep like with like
  - Approaches differ on exactly what is preserved.

SVD

- Enables creation of reduced dimension model
  - Low rank approximation of original matrix
    - Best-fit at that rank (in least-squares sense)

- Motivation:
  - Original matrix: high dimensional, sparse
    - Similarities missed due to word choice, etc
  - Create new projected space
    - More compact, better captures important variation
  - Landauer et al argue identifies underlying “concepts”
    - Across words with related meanings

Document Context

- All models so far:
  - Term x term (or term x relation)
- Alternatively:
  - Term x document
    - Vectors of occurrences (association) in “document”
      - Document can be:
        - Typically: article, essay, etc
        - Also, utterance, dialog act
- Well-known term x document model:
  - Latent Semantic Analysis (LSA)

LSA Document Contexts

- (Deerwester et al, 1990)
- Titles of scientific articles

Example of text data: Titles of Some Technical Memos

c1:	*Human* machine *interface* for ABC *computer* applications
c2:	A *survey* of *user* opinion of *computer system response time*
c3:	The *EPS user interface* management *system*
c4:	*System* and *human system* engineering testing of *EPS*
c5:	Relation of *user* perceived *response time* to error measurement
m1:	The generation of random, binary, ordered *trees*
m2:	The intersection *graph* of paths in *trees*
m3:	*Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4:	*Graph minors*: A *survey*

Document Context Representation

- Term x document:

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Document Context Representation

- Term x document:
  - Corr(human,user) = -0.38; corr(human,minors)=-0.29

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Improved Representation

- Reduced dimension projection:
  - Corr(human,user) = 0.98; corr(human,minors)=-0.83

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.16	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.42
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62

Diverse Applications

- Unsupervised POS tagging
- Word Sense Disambiguation
- Essay Scoring
- Document Retrieval
- Unsupervised Thesaurus Induction
- Ontology/Taxonomy Expansion
- Analogy tests, word tests
- Topic Segmentation

Distributional Similarity for Word Sense Disambiguation

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

Word Representation

- 2nd order representation:

  - Identify words in context of *w*

  - For each x in context of w
    - Compute x's vector representation

  - Compute centroid of those x vector representations

Computing Word Senses

- Compute context vector for each occurrence of word in corpus
- Cluster these context vectors
  - # of clusters = # number of senses
- Cluster centroid represents word sense
- Link to specific sense?
  - Pure unsupervised: no sense tag, just i^{th} sense
  - Some supervision: hand label clusters, or tag training

Disambiguating Instances

- To disambiguate an instance t of w:
  - Compute context vector for the instance
  - Retrieve all senses of w
  - Assign w sense with closest centroid to t

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 and many others. We use reagent injection in molten metal for the…

Industrial Example

Label the First Use of "Plant"

Example Sense Selection for Plant Data

- 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

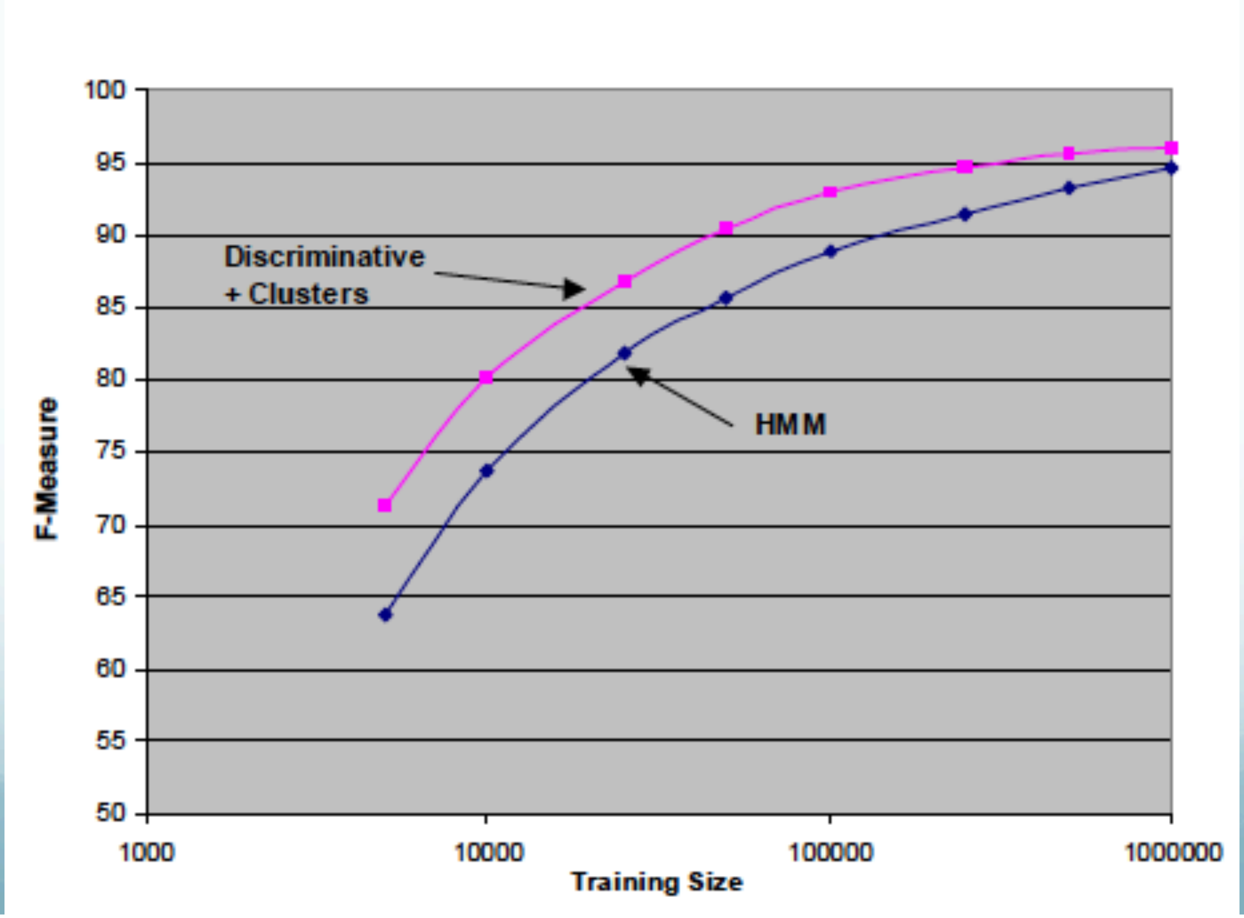
Local Context Clustering

- "Brown" (aka IBM) clustering (1992)
  - Generative model over adjacent words
  - Each w_i has class c_i
  - $\log P(W) = \Sigma_i \log P(w_i|c_i) + \log P(c_i|c_{i-1})$
    - (Familiar??)

  - Greedy clustering
    - Start with each word in own cluster
    - Merge clusters based on log prob of text under model
      - Merge those which maximize $P(W)$

Clustering Impact

- Improves downstream tasks
  - Here Named Entity Recognition vs HMM (Miller et al '04)



Distributional Models

- Upsurge in distributional compositional models
  - Neural network embeddings:
    - Discriminatively trained, low dimensional reps
    - E.g. word2vec
      - Skipgrams etc over large corpora
  - Composition:
    - Methods for combining word vector models
      - Capture phrasal, sentential meanings

Dictionary-Based Approach

- (Simplified) Lesk algorithm
  - “How to tell a pine cone from an ice cream cone”
- Compute ‘signature’ of word senses:
  - Words in gloss and examples in dictionary
- Compute context of word to disambiguate
  - Words in surrounding sentence(s)
- Compare overlap b/t signature and context
- Select sense with highest (non-stopword) overlap

Applying Lesk

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

$bank^1$	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"

- $Bank^1$: 2
- $Bank^2$: 0

Improving Lesk

- Overlap score:
  - All words equally weighted (excluding stopwords)

- Not all words equally informative
  - Overlap with unusual/specific words – better
  - Overlap with common/non-specific words – less good
- Employ corpus weighting:
  - IDF: inverse document frequency
    - $Idf_i = \log(Ndoc/nd_i)$

Thesaurus-Based Similarity

WordNet Taxonomy

- Most widely used English sense resource
- Manually constructed lexical database
  - 3 Tree-structured hierarchies
    - Nouns (117K) , verbs (11K), adjective+adverb (27K)
    - Entries: synonym set, gloss, example use
- Relations between entries:
  - Synonymy: in synset
  - Hypo(per)nym: Isa tree

WordNet

The noun "bass" has 8 senses in WordNet.

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

The adjective "bass" has 1 sense in WordNet.

1. bass^{1}, deep^{6} - (having or denoting a low vocal or instrumental range)
   "a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"

Noun WordNet Relations

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Instance Hypernym	Instance	From instances to their concepts	$Austen^1 \rightarrow author^1$
Instance Hyponym	Has-Instance	From concepts to concept instances	$composer^1 \rightarrow Bach^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Member Holonym	Member-Of	From members to their groups	$copilot^1 \rightarrow crew^1$
Part Meronym	Has-Part	From wholes to parts	$table^2 \rightarrow leg^3$
Part Holonym	Part-Of	From parts to wholes	$course^7 \rightarrow meal^1$
Substance Meronym		From substances to their subparts	$water^1 \rightarrow oxygen^1$
Substance Holonym		From parts of substances to wholes	$gin^1 \rightarrow martini^1$
Antonym		Semantic opposition between lemmas	$leader^1 \Longleftrightarrow follower^1$
Derivationally Related Form		Lemmas w/same morphological root	$destruction^1 \Longleftrightarrow destroy^1$

WordNet Taxonomy

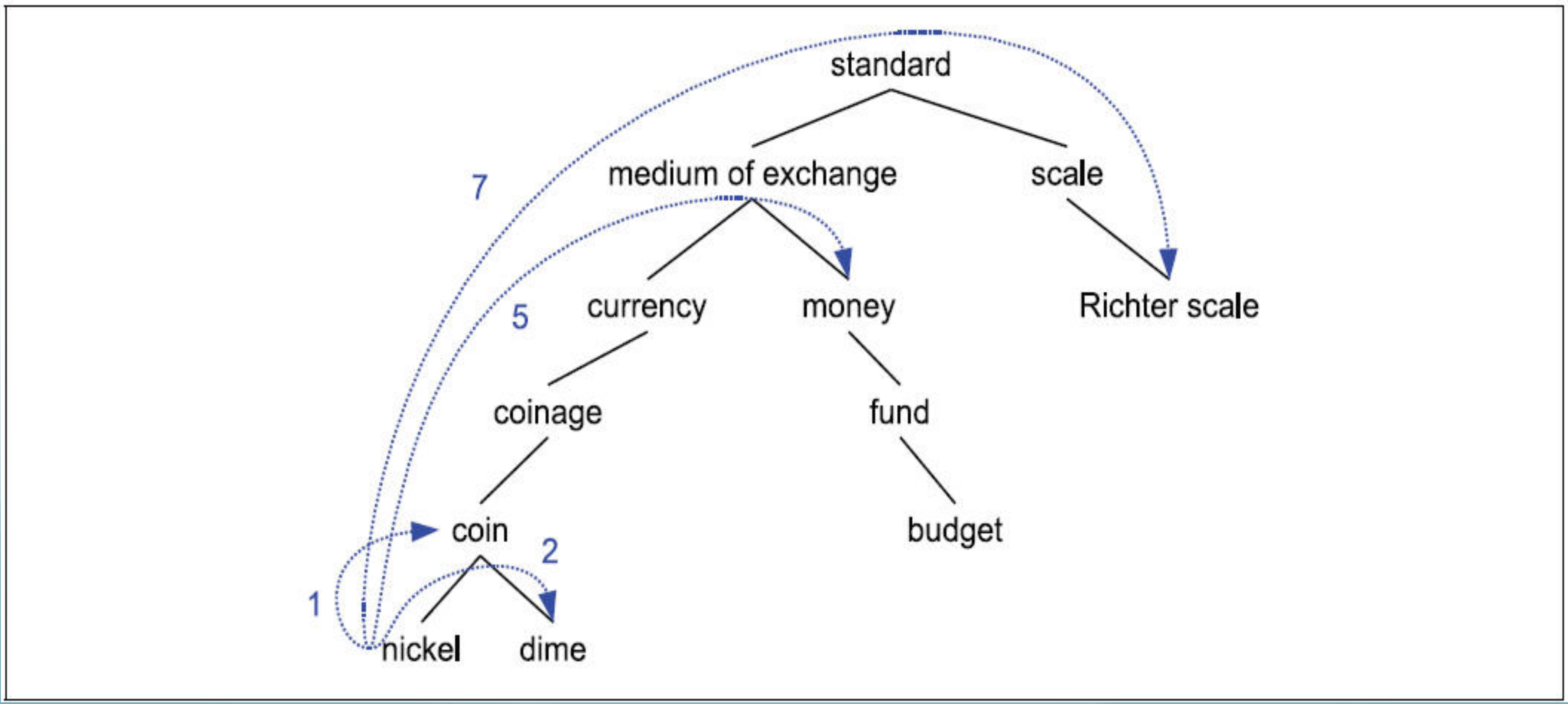
```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity
```

Thesaurus-based Techniques

- Key idea:
  - Shorter path length in thesaurus, smaller semantic dist.
    - Words similar to parents, siblings in tree
      - Further away, less similar
- Pathlength=# edges in shortest route in graph b/t nodes
  - Sim_{path}= -log pathlen(c_1 ,c_2) [Leacock & Chodorow]
- Problem 1:
  - Rarely know which sense, and thus which node
- Solution: assume most similar senses estimate
  - Wordsim(w_1,w_2) = max sim(c_1,c_2)

Path Length

- Path length problem:
  - Links in WordNet not uniform
    - Distance 5: Nickel->Money and Nickel->Standard

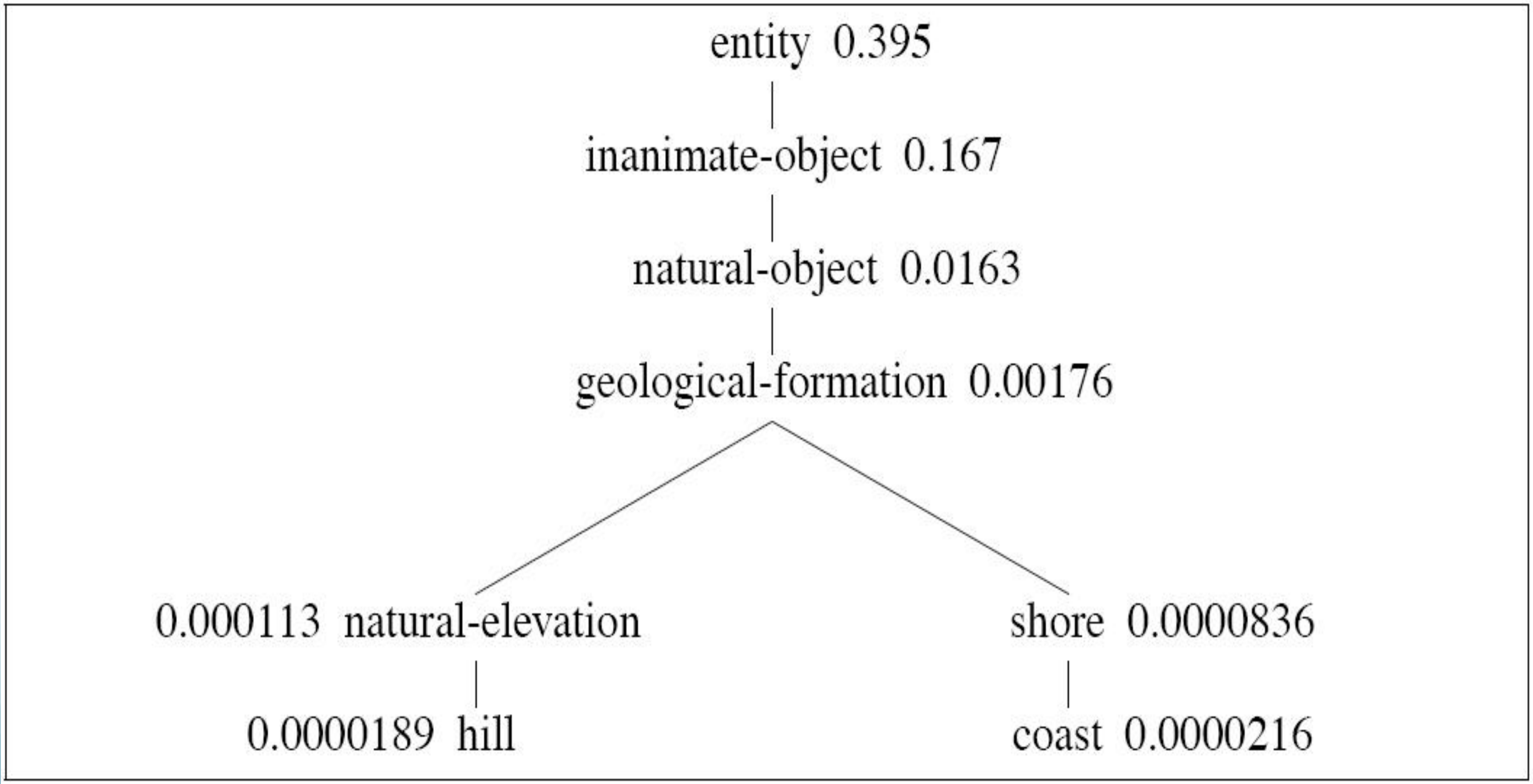


Resnik's Similarity Measure

- Solution 1:
  - Build position-specific similarity measure
  - Not general
- Solution 2:
  - Add corpus information: information-content measure
    - P(c) : probability that a word is instance of concept c
      - Words(c) : words subsumed by concept c; N: words in corpus

$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$

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

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

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