Thesaurus-Based Similarity

Ling571
Deep Processing Techniques for NLP
March 2, 2015

Roadmap

- Lexical Semantics
  - Thesaurus-based Word Sense Disambiguation
    - Taxonomy-based similarity measures
    - Disambiguation strategies
  - Semantics summary
- Discourse:
  - Introduction & Motivation
  - Coherence
  - Co-reference

Previously

- Features for WSD:
  - Collocations, context, POS, syntactic relations
  - Can be exploited in classifiers
- Distributional semantics:
  - Vector representations of word “contexts”
    - Variable-sized windows
    - Dependency-relations
  - Similarity measures
- But, no prior knowledge of senses, sense relations

Exploiting Sense Relations

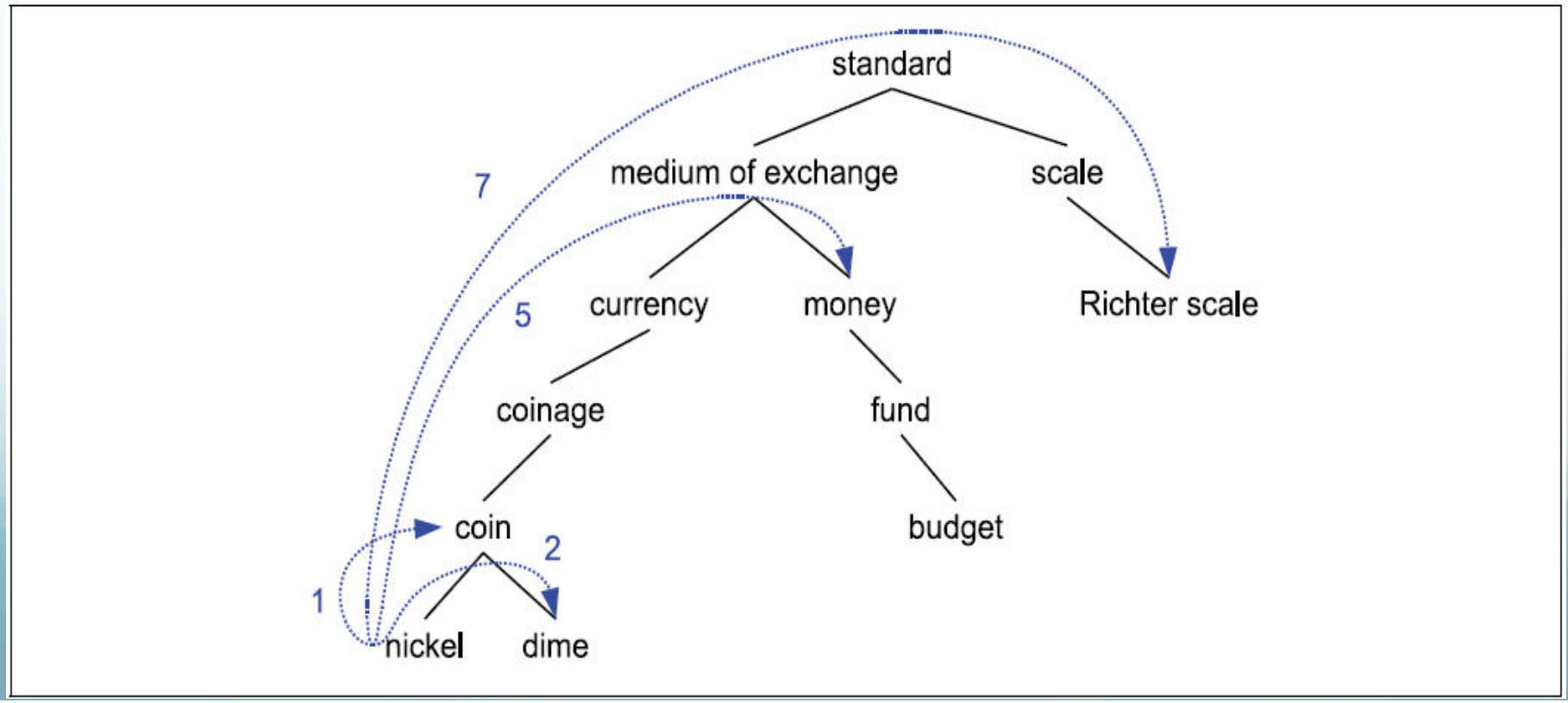
- Distributional models don't use sense resources
- But, we have good ones, e.g.
- WordNet!
  - Also FrameNet, PropBank, etc
- How can we leverage WordNet taxonomy for WSD?

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

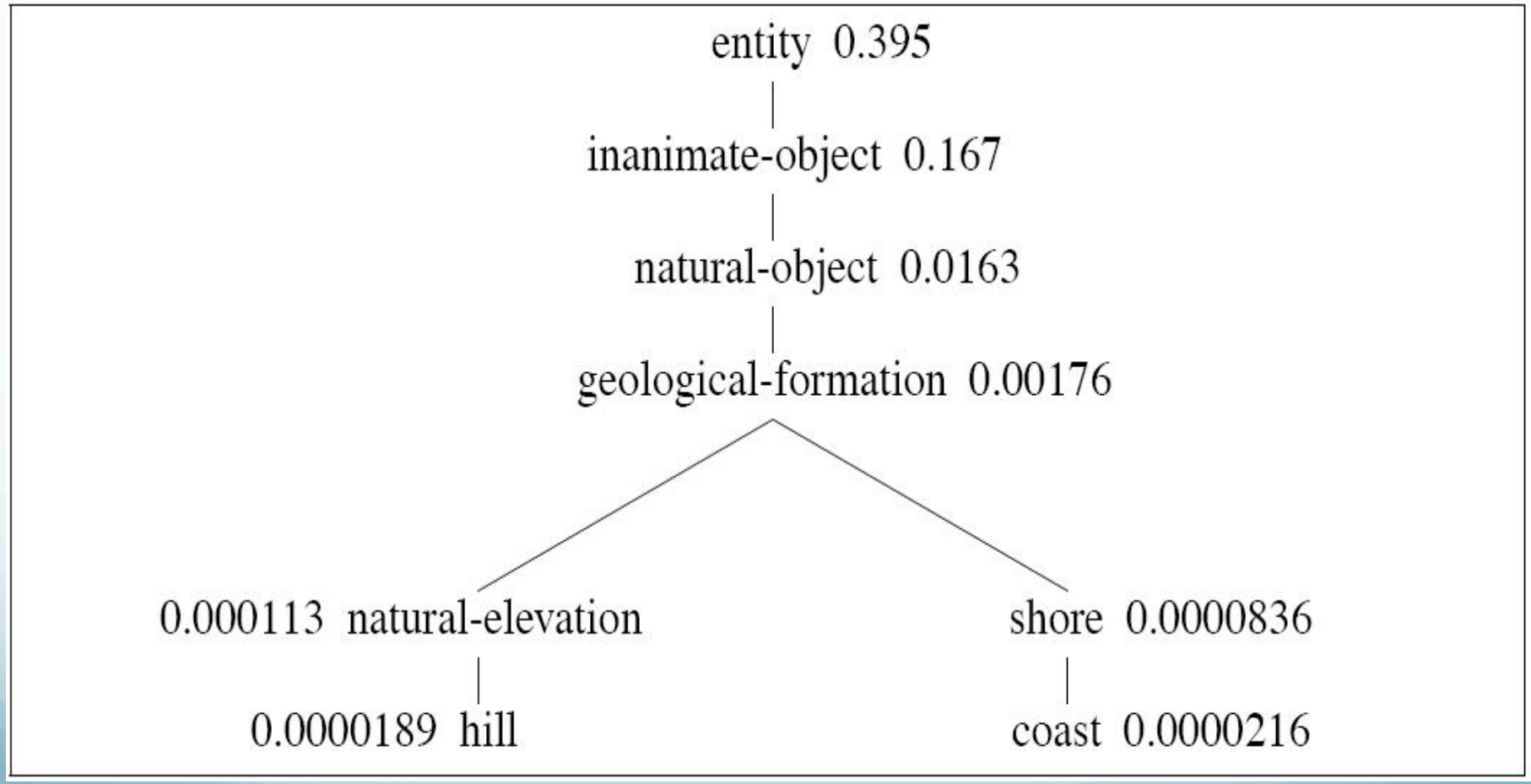


Resnik's Similarity Measure

- 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
0.0000189 hill
shore 0.0000836
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

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

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

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”

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

- 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

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”
  - Selectional Restrictions Override Topic
- No Surface Regularities
  - “The emigration/immigration bill guaranteed passports to all Soviet citizens
  - No Substitute for Understanding

Summary

- Computational Semantics:
  - Deep compositional models yielding full logical form
  - Semantic role labeling capturing who did what to whom
  - Lexical semantics, representing word senses, relations

Computational Models of Discourse

Roadmap

- Discourse
  - Motivation
  - Dimensions of Discourse
  - Coherence & Cohesion
  - Coreference

What is a Discourse?

- Discourse is:
  - Extended span of text
  - Spoken or Written
  - One or more participants
  - Language in Use
  - Goals of participants
    - Processes to produce and interpret

Why Discourse?

- Understanding depends on context
  - Referring expressions: it, that, the screen
  - Word sense: plant
  - Intention: Do you have the time?
- Applications: Discourse in NLP
  - Question-Answering
  - Information Retrieval
  - Summarization
  - Spoken Dialogue
  - Automatic Essay Grading

Reference Resolution

U: Where is A Bug's Life playing in Summit?
S: A Bug's Life is playing at the Summit theater.
U: When is it playing there?
S: It's playing at 2pm, 5pm, and 8pm.
U: I'd like 1 adult and 2 children for the first show.
How much would that cost?

- Knowledge sources:
  - Domain knowledge
  - Discourse knowledge
  - World knowledge

From Carpenter and Chu-Carroll, Tutorial on Spoken Dialogue Systems, ACL '99

Coherence

- *First Union Corp. is continuing to wrestle with severe problems. According to industry insiders at PW, their president, John R. Georgius, is planning to announce his retirement tomorrow.*
- Summary*:*
- *First Union President John R. Georgius is planning to announce his retirement tomorrow.*

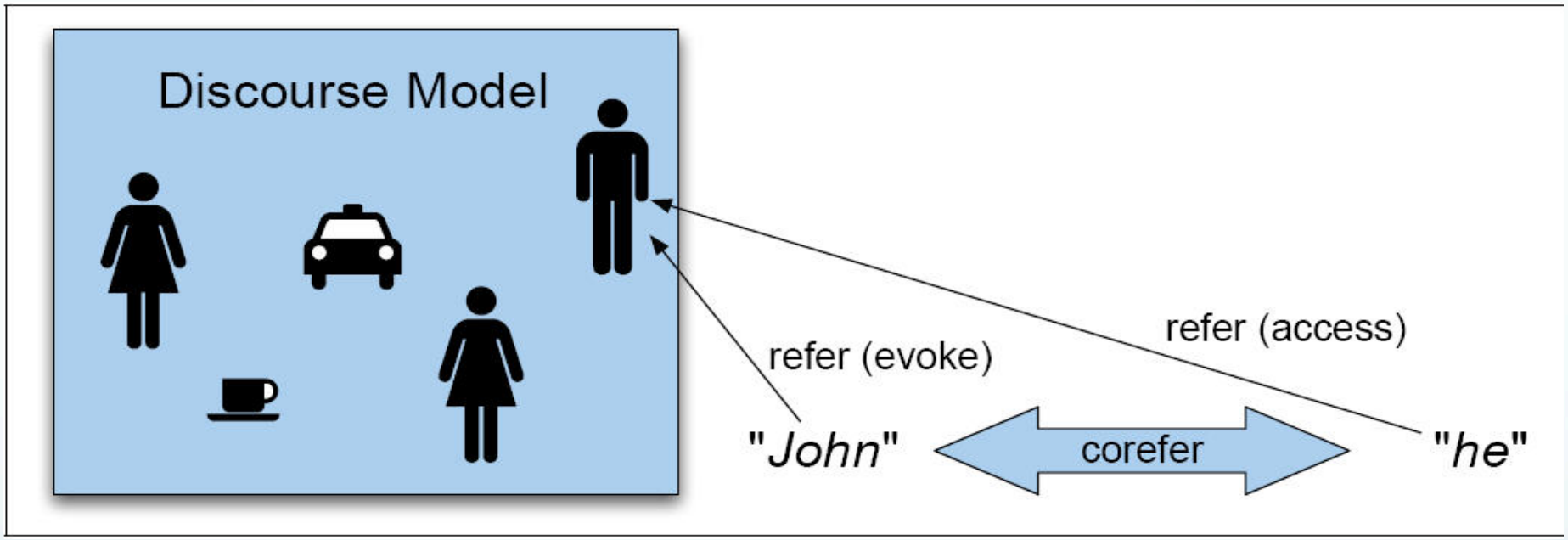
- Inter-sentence coherence relations:
  - Second sentence: main concept (nucleus)
  - First sentence: subsidiary, background

Coherence Relations

- John hid Bill's car keys. He was drunk.
- ?? John hid Bill's car keys. He likes spinach.

- Why odd?
  - No obvious relation between sentences
    - Readers often try to construct relations
- How are first two related?
  - Explanation/cause
- Utterances should have meaningful connection
  - Establish through **coherence relations**

Reference and Model



Reference Resolution

- Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

Coreference resolution:

Find all expressions referring to same entity, ‘corefer’

Colors indicate coreferent sets

Pronominal anaphora resolution:

Find antecedent for given pronoun