Thesaurus-Based Similarity

Ling571
Deep Processing Techniques for NLP
February 29, 2016

Roadmap

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

Previously

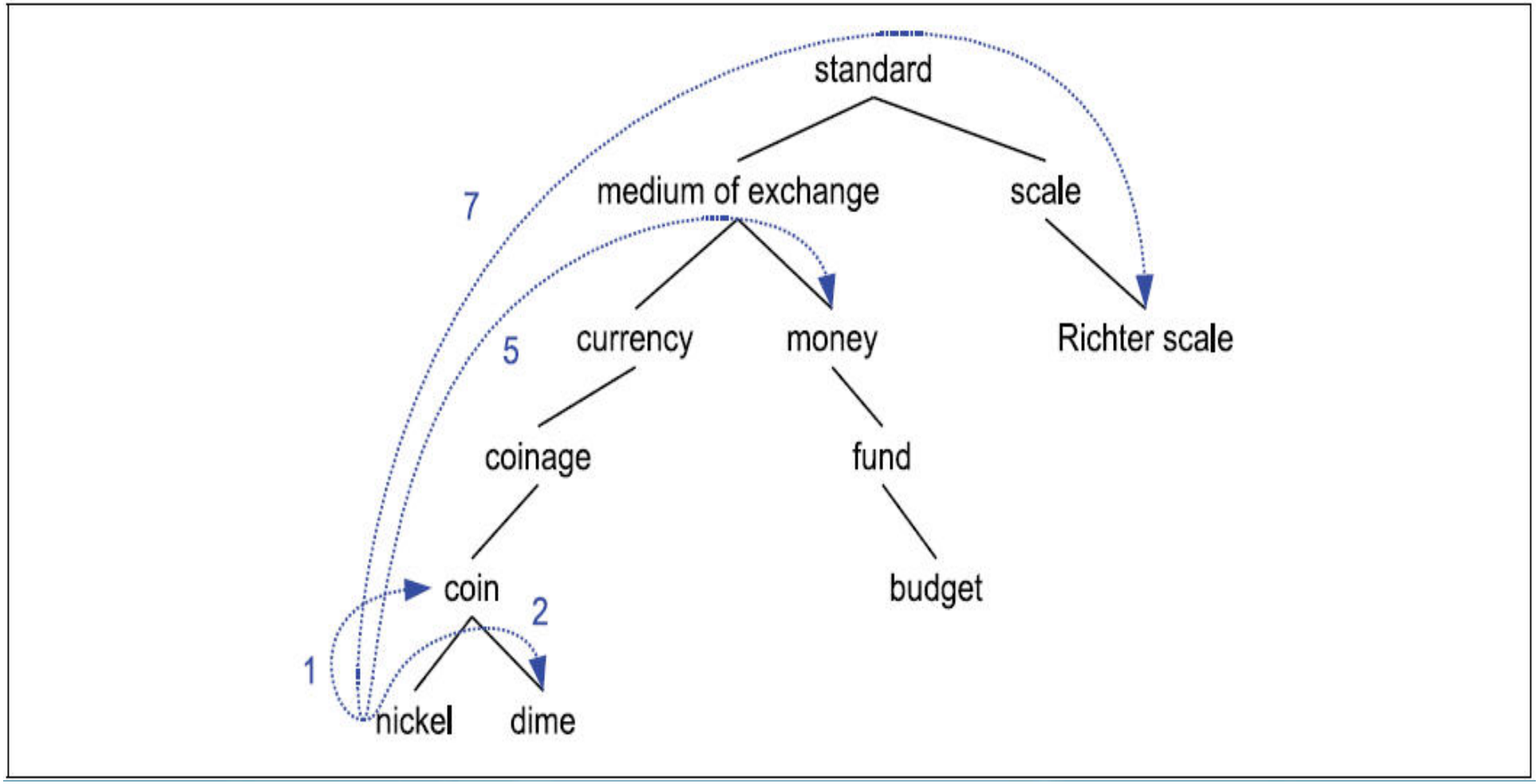
- Features for WSD:
  - Collocations, context, POS, syntactic relations
  - Can be exploited in classifiers
- Distributional semantics:
  - Vector representations of word “contexts”
    - Variable-sized windows
    - Dependency-relations
  - Similarity measures
- But, no prior knowledge of senses, sense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

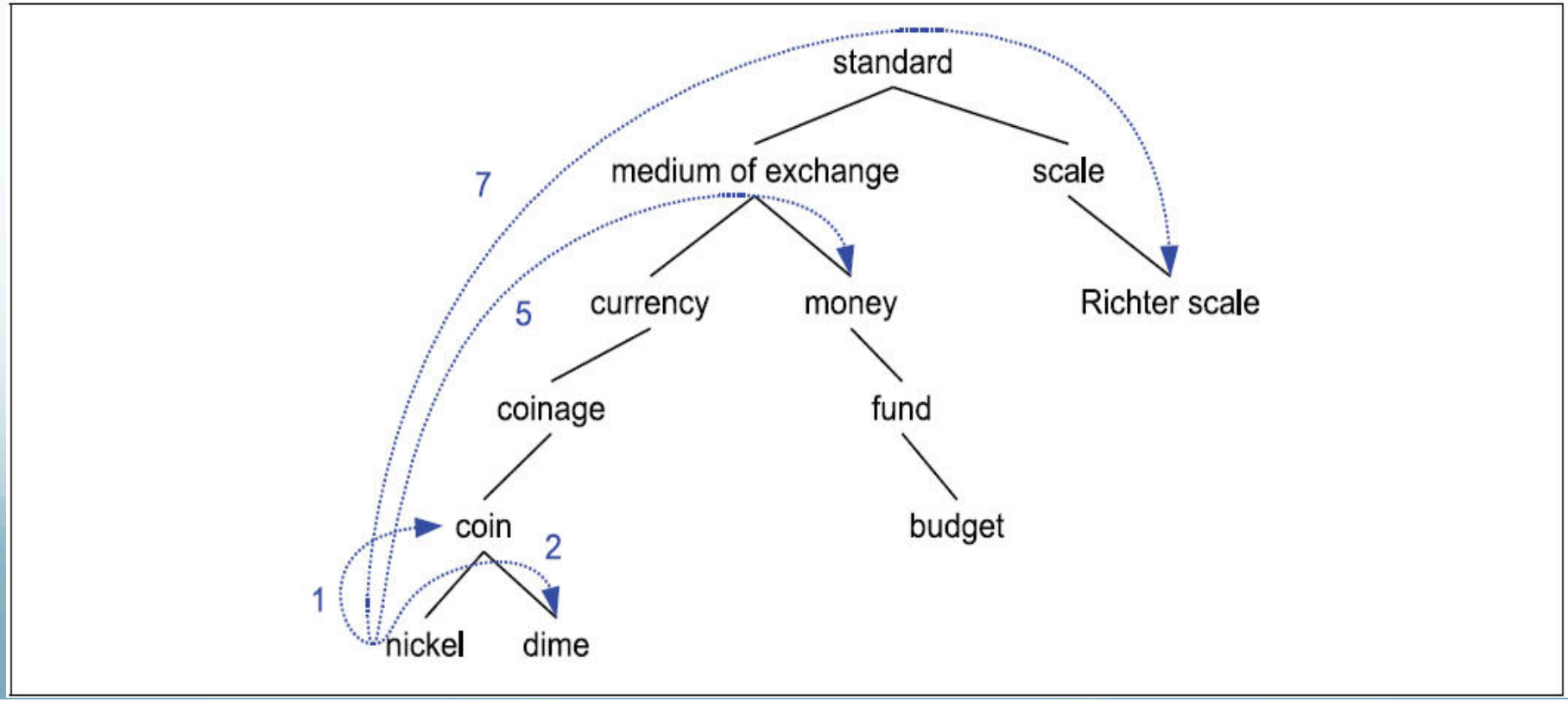
Path Length

- Path length problem:



Path Length

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



Information Content-Based Similarity Measures

- Issues:
  - Word similarity vs sense similarity
    - Assume: sim(w1,w2) = $\max_{si:wi;sj:wj}$ (si,sj)
  - Path steps non-uniform
- Solution:
  - 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

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

Information Content-Based Similarity Measures

- Information content of node:
  - $IC(c) = -\log P(c)$
- Least common subsumer (LCS):
  - Lowest node in hierarchy subsuming 2 nodes
- Similarity measure:
  - $sim_{RESNIK}(c_1,c_2) = -\log P(LCS(c_1,c_2))$

Information Content-Based Similarity Measures

- 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

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

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

Issues:

Underflow => log prob

Sparseness => smoothing

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

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

Different Parameters of Discourse

- Number of participants
  - Multiple participants -> Dialogue
- Modality
  - Spoken vs Written
- Goals
  - Transactional (message passing) vs Interactional (relations,attitudes)
  - Cooperative task-oriented rational interaction

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**

Entity-based Coherence

- *John went to his favorite music store to buy a piano.*
- *He had frequented the store for many years.*
- *He was excited that he could finally buy a piano.*

- VS
  - *John went to his favorite music store to buy a piano.*
  - *It was a store John had frequented for many years.*
  - *He was excited that he could finally buy a piano.*
  - *It was closing just as John arrived.*

- Which is better? Why?
  - 'about' one entity vs two, focuses on it for coherence