Seminar

Recent Developments in Computational Semantics

Word Similarity Measures

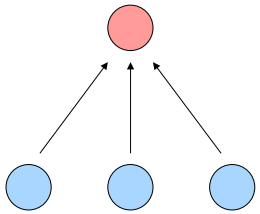
Manfred Pinkal
Saarland University
Summer 2010

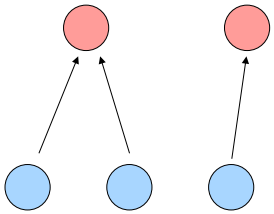
Reading

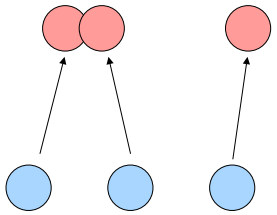
Background Reading:

Jurafsky&Martin, Ch. 20.6+7 (p. 686 – 701)

The Problem: Different words – Same or related senses







WordNet Relations

- Synonymy
- Hyponymy
- Meronymy
- Antonymy
- (+ some additional relations for verbs)

Concept Overlap

Aki Kaurismäki directed his first full-time feature
Aki Kaurismäki directed a film

A car accident occurred yesterday
A vehicle accident occurred yesterday

Several airlines polled saw costs grow more than expected, even after adjusting for inflation
Some companies reported cost increases

WordNet Similarity

- A simple distance measure: Path length

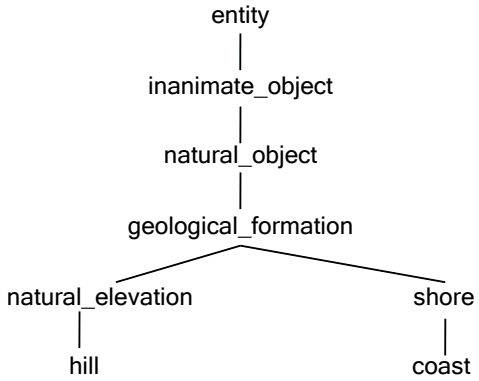
$$dist_{WN} = pathlength(s_1, s_2)$$

- A simple similarity measure: inverse of path length

$$sim_{WN} = \frac{1}{pathlength(s_1, s_2)}$$

- WordNet Similarity measures typically make use of hyponymy only

WordNet Similarity and Information content 2



WordNet Similarity

- Problem 1: Semantic similarity is a relation between word-senses rather than words. In typical applications, we do not have (immediate, reliable) access to word senses
- Standard approach: Define the similarity between w and w' as the similarity between the minimally distant sense pair (s,s') of w and w', respectively.

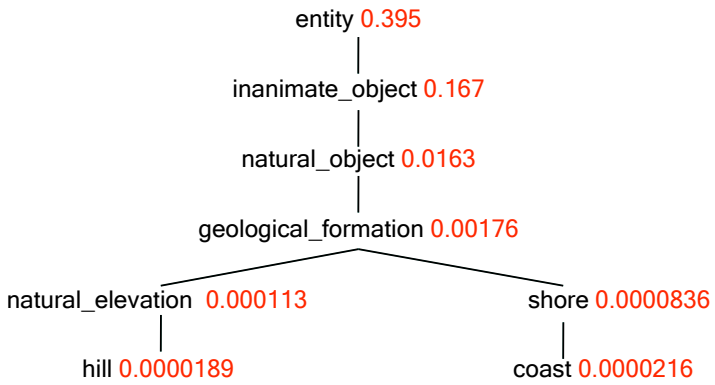
- Problem 2: Absolute pathlength in general is not a fully appropriate measure of semantic distance
- Simple solution: Normalize, e.g., by path length from root to lowest common subsumer/ hypernym.

Measuring Shared Information Content

- Take the lowest common hypernym s of s1 and s2 to represent the shared information between s1 and s2
- Measure the information content of s.
- But how?
- The less frequent a concept is used, the higher its information content. So, first, we compute the instantiation probability of s:
  - *words(s)* is the set of words subsumed by a synset s, i.e.: all words in the concept's synset plus all words in synsets which are hyponyms to s.
  - Instantiation probability of synset:

$$P(s) = \frac{\sum_{w \in words(s)} count(w)}{corpus\_size}$$

WordNet Similarity and Information content 2



Information Content

- *words(c)* is the set of words subsumed by a synset s, i.e.: all words in the concept's synset plus all words in synsets which are hyponyms to s.
- Instantiation probability of synset:

$$P(s) = \frac{\sum_{w \in words(c)} count(w)}{corpus\_size}$$

- Information content of synset:

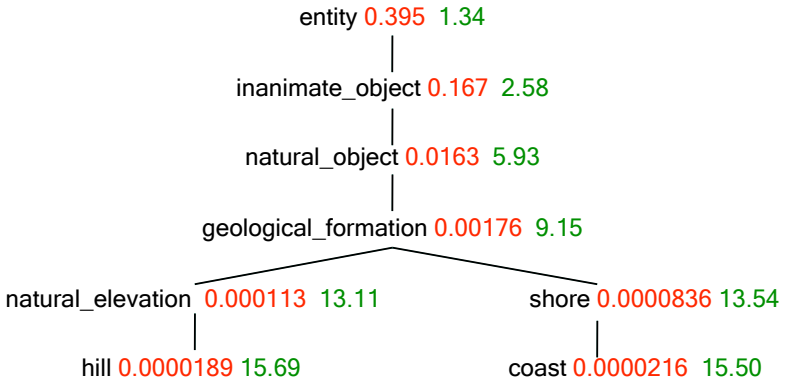
$$IC(s) = -\log P(s)$$

WordNet Similarity and Information content 1

- First approximation: To compute similarity between A and B, measure the amount of information shared by A and B.

$$sim_{resnik}(s_1, s_2) = -\log P(LCS(s_1, s_2))$$

WordNet Similarity and Information content 2

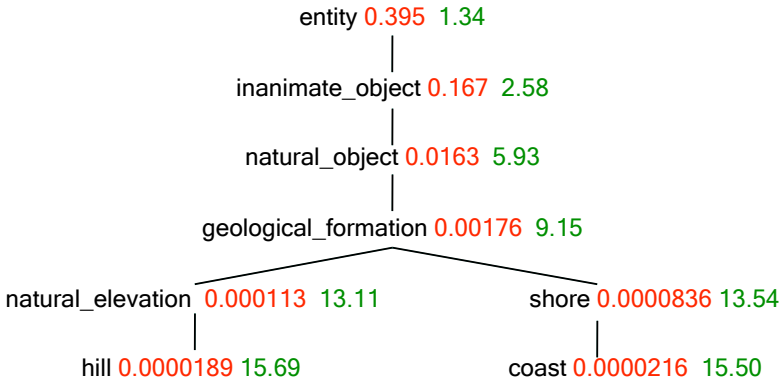


WordNet Similarity and Information content 2

- Lin's WordNet similarity measure (Lin 1997): Similarity between A and B is the ratio between
  - the amount of information shared by A and B, and
  - the cumulative information content of A and B.

$$sim_{lin}(s_1, s_2) = \frac{2 * \log P(LCS(s_1, s_2))}{\log P(s_1) + \log P(s_2)}$$

WordNet Similarity and Information content 2



WordNet Similarity and Information content 3

- Jiang-Conrath distance (Jiang&Conrath 1997): Distance between A and B is the difference between
  - the amount of information shared by A and B, and
  - the cumulative information content of A and B.

$$dist_{JC}(s_1,s_2) = 2 * \log P(LCS(s_1,s_2)) - (\log P(s_1) + \log P(s_2))$$

- Jiang-Conrath similarity: Negative reciprocal distance:

$$sim_{JC}(s_1,s_2) = -\frac{1}{dist_{JC}(s_1,s_2)}$$

Lesk Measure

Yet another resource-based similarity measure:

Based on phrase overlap between glosses.

Best performing measures are Jiang-Conrath and an extended Lesk variant.

Limitations of lexicon-based similarity

- Limited coverage of WordNet
  - Missing words
  - Varying depth of hierarchy
  - Fewer hyponymy relations for verbs, none for adjectives
  - No (or very few) hyponymy links between nouns and verbs
- Limited adaptability
  - new domains (special terminology, constrained semantics)
  - new developments (neologisms, semantic change)