Seminar

Recent Developments in Computational Semantics

Word Similarity Measures II

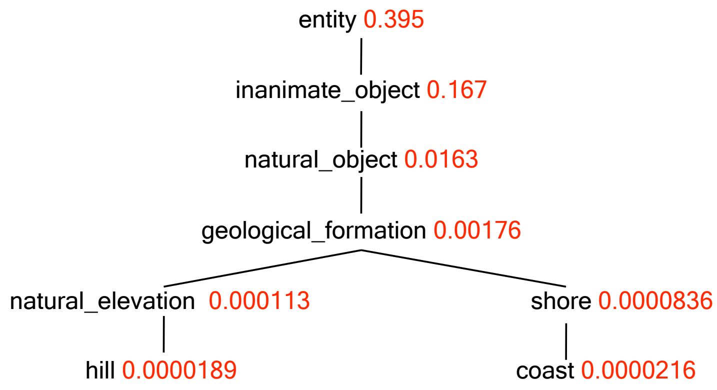
Manfred Pinkal
Saarland University
Summer 2010

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

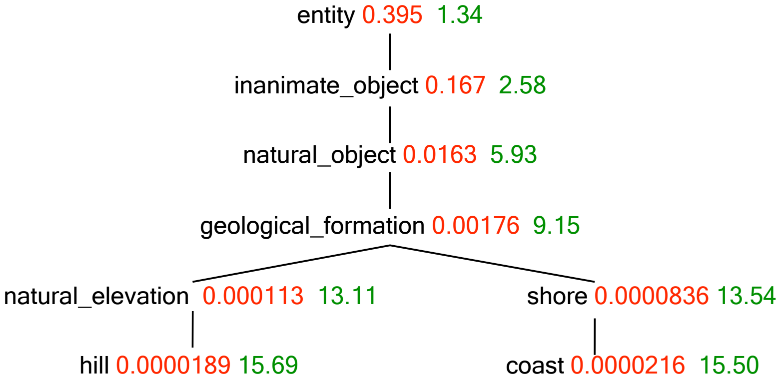
- *words(c)* is the set of words subsumed by a synset s, i.e.: all words in the concept's synset plus all words in synsets which are hyponyms to s.
- Instantiation probability of synset:

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

- Information content of synset:

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

WordNet Similarity and Information content 2

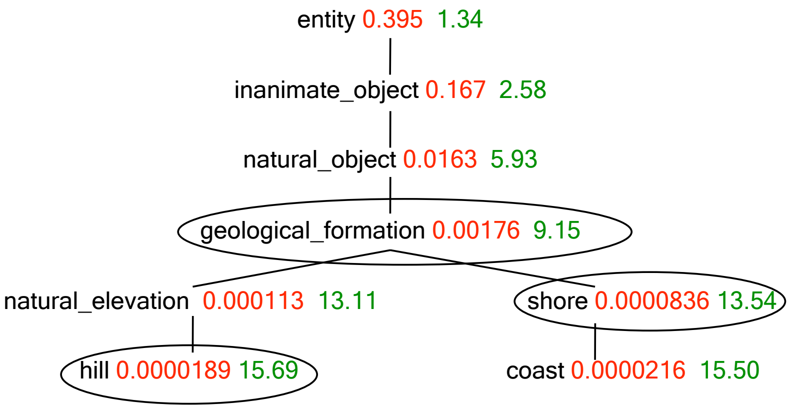


WordNet Similarity and Information content 2

- 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 measures

- 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)

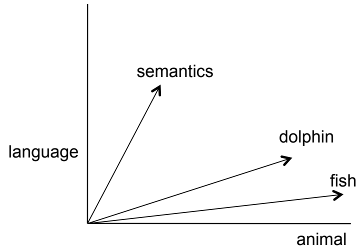
Co-ocurrence vectors

- Distributional hypothesis:
  Two words are semantically similar to the extent that they occur in similar contexts.
- Context of a word w:
  - A window containing n (5, 10, 50, ...) words before and after an occurrence of w.
- Features used for the description of contexts are context words
- Representation of a word w's typical context (distributional "meaning representation" of w):
  - Count the number of occurrences of all all content words across all contexts of w (in a corpus).
  - Take the function from the considered context words to occurrency frequencies as context representation for w.
  - This is a vector in a multi-dimensional space (the "word space").

Simple Example

- Frequencies of 'animal' and 'language' in the context of 'dolphin', 'fish', and 'semantics'.

	dolphin	semantics	fish
animal	55	15	70
language	15	45	5



- The table and its graphical representation indicate the affinity of 'dolphin' and 'fish' to the domains of zoology, and of 'semantics' to language.
- They also indicate that 'dolphin' and 'fish' are more similar to each other than to 'semantics'.

Measuring Similarity

- One standard measure for distributional similarity is cosine:

- Cosine is 1, if vectors have identical directions ($cos(0^0)=1$), it is 0, if vectors are orthogonal ($cos(90^0)=0$).
- General definition:

$$sim\ (x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

- In our example:

$$sim(semantics, dolphin) = 0.55$$
$$sim(semantics, fish) = 0.38$$
$$sim(fish, dolphin) = 0.98$$

Wait a minute ...

- Strong distributional hypothesis (Schütze 1998):
  Two words are semantically similar to the extent that they occur in similar contexts.
- A more cautious classical formulation (Harris 1968):
  The meaning of entities ... is related to the restriction of combinations of these entities relative to other entities.
- Distributional similarity and semantic similarity cannot be simply identified:
  - Distributional similarity is measured on words, not on word senses
  - Distributional similarity: Semantic similarity or semantic relatedness?
  - How can appropriateness of similarity measures be evaluated?

Similarity and Relatedness

- Similarity: Quasi-synonymy, information-preserving substitutability in context
  - car - automobile, walk - stroll, fast - quick

- Relatedness: A much more general kind of semantic proximity, comprising topical relatedness, collocations, meronymy, antonymy:
  - car - drive - highway - engine - flat tire
  - red - blue
  - short - long

- http://clg.wlv.ac.uk/demos/similarity/

Evaluation of Similarity Measures

- Association tests with human subjects
- Similarity scores assigned by humans
- Evaluating against a gold-standard thesaurus

- End-to-end evaluations in NLP tasks (e.g., WSD)

Questions to be asked

- What kind of context is taken into account? What are the dimensions of the feature vector?
- How is the association between words and context features measured?
- How is vector similarity defined?

Answers for the simple model

1. What kind of context is taken into account? What are the dimensions of the feature vector
   Context window of size n, dimensions are content words
2. How is the association between words and context features measured?
   Frequency of context words for a given w
3. How is similarity defined?
   Cosine

(1) Context and Feature Space

Options:

Context windows and word space

Syntactically structured context, syntax-sensitive feature space (Lin 1998):

- Context is the syntactically analysed sentence.
- Syntactic analysis done by a dependency parser.
- Structural information given in terms of dependency triples (w,r,w').

Lin's Example

- Dependency triples for

  I have a brown dog

  (have subj I), (I subj-of have), (dog obj-of have), (dog adj-mod brown), (brown adj-mod-of dog), (dog det a), (a det-of dog)

- ***Frequency counts for "cell"***

||cell, pobj-of, in||=159
||cell, pobj-of, inside||=16
||cell, pobj-of, into||=30
......
||cell, nmod-of, abnormality||=3
||cell, nmod-of, anemia||=8
||cell, nmod-of, architecture||=1
......
||cell, obj-of, attack||=6
||cell, obj-of, bludgeon||=1
||cell, obj-of, call||=11

(2) Association with context

Options for feature values:

- (Relative) frequencies, probabilities (w' occurs 3785 times/ with a frequency of 0.217 in the context of w)
- Binary values (w' occurs/ doesn't occur in the context of w)
- (Pointwise) Mutual Information (PMI)

PMI

- Measure of the co-occurrence of two events exceeding random probability
  - 0, if randomly distributed,
  - positive/negative, if positively/negatively correlated

$$I(x,y) = \log \frac{P(x,y)}{P(x) * P(y)}$$

- PMI-based co-occurrence values in a BOW setting:
  Let $f_{w'}$ be the feature "w' occurring as a context word".
  Then the PMI-based value of $f_{w'}$ for w is:

$$f_{w'}(w) = I(w,w') = \log \frac{P(w,w')}{P(w) * P(w')}$$

PMI for Dependency Triples

$$I(x,y) = \log \frac{P(x,y)}{P(x) * P(y)}$$

$$f_{r,w'}(w) = I(w,r,w') = \log \frac{P(w,r,w')}{P(w) * P(w \mid r) * P(w' \mid r)}$$

$$P_{MLE}(B) = \frac{\|*,r,*\|}{\|*,*,*\|}, \quad P_{MLE}(A|B) = \frac{\|w,r,*\|}{\|*,r,*\|}, \quad P_{MLE}(C|B) = \frac{\|*,r,w'\|}{\|*,r,*\|}$$

$$\begin{aligned} & I(w,r,w') \\ & = -\log(P_{MLE}(B) P_{MLE}(A|B) P_{MLE}(C|B)) \\ & \quad -(-\log P_{MLE}(A,B,C)) \\ & = \log \frac{\|w,r,w\| \times \|*,r,*\|}{\|w,r,*\| \times \|*,r,w'\|} \end{aligned}$$

(3) Similarity Measure

$$\frac{\sum_{(r,w) \in T(w_1) \cap T(w_2)} (I(w_1,r,w) + I(w_2,r,w))}{\sum_{(r,w) \in T(w_1)} I(w_1,r,w) + \sum_{(r,w) \in T(w_2)} I(w_2,r,w)}$$

Comparison of similarity measures

- Coverage and adaptability
  - Wide coverage, easy adaptability of unsupervised distributional methods (provided that raw corpus data are available)
  - In part better precision of WN-based measures

- What do similarity measures express?
  - BOW models measure unspecific relatedness, including topical relatedness
  - Syntax-sensitive models measure similarity in the sense of (semi-)equivalence or substitutibility
  - All distributional measures have difficulties in excluding antonymies (detecting opposite polarity)