Sémantique lexicale computationnelle

Traitement sémantique automatique des langues naturelles

Pascal Denis

INRIA

pascal.denis@inria.fr

(Slides réalisés à partir du cours de Dan Jurafsky au LSA 2008 et de celui de Katrin Erk à UT Austin 2007)

Master en Linguistique Informatique
Université Paris 7

Année 2009-2010

Outline

- Introduction
- Lexical semantics basics
- Online resources: WordNet
- Computational lexical semantics
  - **Word Sense Disambiguation (WSD)**
  - Semantic Role Labeling (SRL)
  - **Word similarity**
  - Lexical Acquisition

Introduction

Introduction

- So far, we've learned to compute and interpret semantic representations
  - A man bought a donkey.
  - $\exists x, y, t\ [man'(x) \wedge donkey'(y) \wedge buy'(x, y, t) \wedge t < now]$
- Nice, but we still don't have a full understanding of this sentence
  - what *man'*, *donkey'*, and *buy'* actually mean?
  - no treatment of lexical ambiguity: e.g., John went to the bank
  - no way to draw inferences: A man bought an animal, A man now owns a donkey, ...

Three Perspectives on Meaning

- Lexical Semantics
  - The meanings of individual words
- Compositional Semantics
  - How those meanings combine to make meanings for individual sentences/clauses or utterances
- Discourse Semantics/Pragmatics
  - How sentence/clause meanings combine with each other and with other facts about various kinds of context to make meanings for a text or discourse
  - Dialog or Conversation is often lumped together with Discourse

Lexical Semantics Basics

Outline

- What's a word (for lexical semantics)?
- How to represent word meaning?
- Lexical ambiguity
    - homonymy
    - polysemy
- Lexical relations
    - Synonymy vs. Antonymy
    - Hypernomy vs. Hyponomy
    - Meronomy vs. Holonymy

What's a word?

- What a word is varies according to uses: tokens, stems, lemmas,...
- For lexical semantics, the unit we want is the lemma
  - Singular form for nouns: *carpet* is the lemma for *carpets*
  - Infinitive form for verbs: *faire* is the lemma for *feras*
- Lemmatization is the process of mapping a word form to a lemma (not always deterministic: e.g. *found*)
- Lexeme: An abstract pairing of a lemma with a single meaning representation
- Lexicon: A collection of lexemes

How to represent word meaning?

- In theoretical linguistics:
  - Definitional: necessary and sufficient conditions
    - *bird(x) iff animal(x) & has-wings(x) & ...*
  - Semantic primitives
    - *hen*: `+chicken, +adult, +female`
    - KILL(x,y) <=> CAUSE(x,BECOME(NOT(ALIVE(y))))

- In computational linguistics:
  - Relational: semantic network (e.g., Wordnet)
    - words in relation to other words (IS-A, HAS-A, ...)
  - Associative/Distributional
    - similar meanings <=> similar contexts

Lexical ambiguity

- Some lemmas have multiple different meanings:
  - "Instead, a bank can hold the investments in a custodial account in the client's name"
  - "But as agriculture burgeons on the east bank, the river will shrink even more"
- Thus, bank here has two senses:
  - *bank1:* financial institution
  - *bank2:* slop along river
- Most non-rare words have different meanings
- The task of mapping a word to its correct sense is called Word Sense Disambiguation (or WSD)

Types of lexical ambiguities

- Homonymy
  - Accidental: coincidence that the same lemma has those different senses, no relation between senses (doesn't hold across languages)
  - Examples: *bank*, *pitcher*, ...
  - Homonymy vs. homography vs. homophony
- Polysemy
  - Systematic: the relation between the senses
    other words (tends to hold across languages)
  - Examples: *bank*, *chicken*, *plum*, ...
  - Typical relations: BUILDING-ORGANIZATION, ANIMAL-MEAT, TREE-FRUIT

Homonymes:
glace -> ice cream/mirror
batterie -> drums/battery
mousse, tour -> masc/fem

Homographes: un as, tu as

Homophones:
maitre/metre/mettre
paire, pere; mere/mer/maire
Sot, saut, sceau et seau
vers, ver, verre, vert et vair
mer, mère et maire

Lexical ambiguity test

- Consider examples of the word *serve*:
  - *Which flights serve breakfast?*
  - *Does America West serve Philadelphia?*
- The zeugma/copredication test:
  - *?Does United serve breakfast and San Jose?*
- This test works better for homonymy than polysemy:
  - *John works for the bank across the street.*

Synonyms

- Words have the same meaning in some or all contexts:
  - big / large
  - automobile / car
  - vomit / throw up
  - water / H_20
- Two lexemes are synonyms if they can be successfully substituted for each other in all situations
  - If so they have the same propositional meaning
- Few or no examples of perfect synonymy:
  - Why should there be?
  - Sensitivity to register, genre, ...

Relation between senses rather than words

- Consider the words *big* and *large*
- Are they synonyms?
  - How big is that plane?
  - Would I be flying on a large or small plane?
- How about here:
  - Miss Nelson, for instance, became a kind of big sister to Benjamin.
  - ?Miss Nelson, for instance, became a kind of large sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up

Antonyms

- Senses that are opposites with respect to one feature of their meaning
- Otherwise, they are very similar!
  - dark / light
  - short / long
  - hot / cold
  - up / down
  - in / out
- More formally: antonyms define a binary opposition (pretty/ugly) or at opposite ends of a scale (long/short, fast/slow) or "reversives" (rise/fall, up/down)

Hyponymy

- One sense is a hyponym of another if the first sense is more specific, denoting a subclass of the other
  - *car* is a hyponym of *vehicle*
  - *dog* is a hyponym of *animal*
  - *mango* is a hyponym of *fruit*
- Conversely
  - *vehicle* is a hypernym/superordinate of *car*
  - *animal* is a hypernym of *dog*
  - *fruit* is a hypernym of *mango*

Hypernymy more formally

- Extensional:
  - The class denoted by the superordinate
  - extensionally includes the class denoted by the hyponym
- Entailment:
  - A sense A is a hyponym of sense B if being an A entails being a B
- Hyponymy is usually transitive
  - A hypo B and B hypo C entails A hypo C

Other lexical relations

- Meronymy-holonymy: senses are in a part-whole relationship
  - *wheel* is a meronym of *car*
  - *car* is a holonym of *wheel*
- ...

Wordnet

What is Wordnet?

- Hierarchical lexical database for English
- Developed by lexicographers and psycholinguists at Princeton since 1985 (hugely expensive): http://wordnet.princeton.edu
- Basic unit: **synset** (i.e., list of near-synonyms).
- Each entry contains synset, definition, examples, and links to related synsets.
- WN encodes lexical relations: syno-/antonymy, hyper-/hyponymy, mero-/holonymy, etc.
- Wordnets being for other languages (Czech, German, French, . . .), both within and outside the EuroWordnet project (e.g., WOLF).
- WN is heavily used for tasks related to Computational Semantics.

WordNet's coverage

Category	Unique Forms
Noun	117,097
Verb	11,488
Adjective	22,141
Adverb	4,601

Format of Wordnet Entries

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"*

WordNet Noun Relations

Relation	Also called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	$breakfast^1 \rightarrow meal^1$
Hyponym	Subordinate	From concepts to subtypes	$meal^1 \rightarrow lunch^1$
Member Meronym	Has-Member	From groups to their members	$faculty^2 \rightarrow professor^1$
Has-Instance		From concepts to instances of the concept	$composer^1 \rightarrow Bach^1$
Instance		From instances to their concepts	$Austen^1 \rightarrow author^1$
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	$course^7 \rightarrow meal^1$
Antonym		Opposites	$leader^1 \rightarrow follower^1$

WordNet Verb Relations

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From a verb (event) to a specific manner elaboration of that verb	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Opposites	$increase^1 \Longleftrightarrow decrease^1$

WordNet Hierarchies

```
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

Sense 7
bass --
(the member with the lowest range of a family of
musical instruments)
=> musical instrument, instrument
   => device
      => instrumentality, instrumentation
         => artifact, artefact
            => whole, unit
               => object, physical object
                  => physical entity
                     => entity
```

How is "sense" defined in WordNet?

- The set of near-synonyms for a WordNet sense is called a synset (synonym set); it's their version of a sense or a concept
- Example: *chump* as a noun to mean
  - {chump^1, fool^2, gull^1, mark^9, patsy^1, fall guy^1, sucker^1, soft touch^1, mug^2}
- Each of these senses share this same gloss
- Thus for WordNet, the meaning of this sense of *chump* is this list.

Wordnet: hyponyms

- (18)S: (n) **beer** (a general name for alcoholic beverages made by fermenting a cereal (or mixture of cereals) flavored with hops)
  - *direct hyponym* / ***full hyponym***
    - S: (n) draft beer, draught beer (beer drawn from a keg)
    - S: (n) suds (a dysphemism for beer (especially for lager that effervesces))
    - S: (n) lager, lager beer (a general term for beer made with bottom fermenting yeast (usually by decoction mashing); originally it was brewed in March or April and matured until September)
      - S: (n) Munich beer, Munchener (a dark lager produced in Munich since the 10th century; has a distinctive taste of malt)
      - S: (n) bock, bock beer (a very strong lager traditionally brewed in the fall and aged through the winter for consumption in the spring)
      - S: (n) light beer (lager with reduced alcohol content)
      - S: (n) Oktoberfest, Octoberfest (a strong lager made originally in Germany for the Oktoberfest celebration; sweet and copper-colored)
      - S: (n) Pilsner, Pilsener (a pale lager with strong flavor of hops; first brewed in the Bohemian town of Pilsen)
      - S: (n) malt, malt liquor (a lager of high alcohol content; by law it is considered too alcoholic to be sold as lager or beer)
    - S: (n) ale (a general name for beer made with a top fermenting yeast; in some of the United States an ale is (by law) a brew of more than 4% alcohol by volume)
      - S: (n) Weissbier, white beer, wheat beer (a general name for beers made from wheat by top fermentation; usually very pale and cloudy and effervescent)
        - S: (n) Weizenbier (a general name in southern Germany for wheat beers)
        - S: (n) Weizenbock (a German wheat beer of bock strength)
      - S: (n) bitter (English term for a dry sharp-tasting ale with strong flavor of hops (usually on draft))
      - S: (n) Burton (a strong dark English ale)
      - S: (n) pale ale (an amber colored ale brewed with pale malts; similar to bitter but drier and lighter)
      - S: (n) porter, porter's beer (a very dark sweet ale brewed from roasted unmalted barley)
      - S: (n) stout (a strong very dark heavy-bodied ale made from pale malt and roasted unmalted barley and (often) caramel malt with hops)
        - S: (n) Guinness (a kind of bitter stout)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*

Wordnet: hypernyms

- (18)S: (n) **beer** (a general name for alcoholic beverages made by fermenting a cereal (or mixture of cereals) flavored with hops)
  - *direct hyponym* / *full hyponym*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) brew, brewage (drink made by steeping and boiling and fermenting rather than distilling)
      - S: (n) alcohol, alcoholic drink, alcoholic beverage, intoxicant, inebriant (a liquor or brew containing alcohol as the active agent) *"alcohol (or drink) ruined him"*
        - S: (n) beverage, drink, drinkable, potable (any liquid suitable for drinking) *"may I take your beverage order?"*
          - S: (n) food, nutrient (any substance that can be metabolized by an animal to give energy and build tissue)
            - S: (n) substance (a particular kind or species of matter with uniform properties) *"shigella is one of the most toxic substances known to man"*
              - S: (n) matter (that which has mass and occupies space) *"physicists study both the nature of matter and the forces which govern it"*
                - S: (n) physical entity (an entity that has physical existence)
                  - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
          - S: (n) liquid (a substance that is liquid at room temperature and pressure)
            - S: (n) fluid (a substance that is fluid at room temperature and pressure)
              - S: (n) substance (the real physical matter of which a person or thing consists) *"DNA is the substance of our genes"*
                - S: (n) matter (that which has mass and occupies space) *"physicists study both the nature of matter and the forces which govern it"*
                  - S: (n) physical entity (an entity that has physical existence)
                    - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
                - S: (n) part, portion, component part, component, constituent (something determined in relation to something that includes it) *"he wanted to feel a part of something bigger than himself"; "I read a portion of the manuscript"; "the smaller component is hard to reach"; "the animal constituent of plankton"*
                  - S: (n) relation (an abstraction belonging to or characteristic of two entities or parts together)

Wordnet: other relations

- (163)S: (n) **night**, nighttime, dark (the time after sunset and before sunrise while it is dark outside)
  - *direct hyponym* / *full hyponym*
  - *part meronym*
    - S: (n) evening (the early part of night (from dinner until bedtime) spent in a special way) *"an evening at the opera"*
    - S: (n) late-night hour (the latter part of night)
    - S: (n) midnight (12 o'clock at night; the middle of the night) *"young children should not be allowed to stay up until midnight"*
    - S: (n) small hours (the hours just after midnight)
    - S: (n) lights-out (a prescribed bedtime)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *part holonym*
    - S: (n) day, twenty-four hours, twenty-four hour period, 24-hour interval, solar day, mean solar day (time for Earth to make a complete rotation on its axis) *"two days later they left"; "they put on two performances every day"; "there are 30,000 passengers per day"*
  - *antonym*
    - W: (n) day [Opposed to: night] (the time after sunrise and before sunset while it is light outside) *"the dawn turned night into day"; "it is easier to make the repairs in the daytime"*
  - ***derivationally related form***
    - W: (adj) nightly [Related to: night] (happening every night) *"nightly television now goes on until 3:00 or 4:00 a.m."*

Wordnet in NLTK

- Documentation: http://nltk.googlecode.com/svn/trunk/doc/howto/wordnet.html
- WordNet can be imported like this:
  - >>> from nltk.corpus import wordnet as wn
- Getting the synsets:
  - >>> wn.synsets('dog')
  - [Synset('dog.n.01'), Synset('frump.n.01'), Synset('dog.n.03'), Synset('cad.n.01'),Synset('frank.n.02'), Synset('pawl.n.01'), Synset('andiron.n.01'), Synset('chase.v.01')]
  - >>> wn.synsets('dog', pos=wn.VERB)
  - [Synset('chase.v.01')]

Wordnet in NLTK

- Accessing the different parts of a WN entry:
  - >>> wn.synset('dog.n.01').definition
  - 'a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds'
  - >>> wn.synset('dog.n.01').examples
  - ['the dog barked all night']
  - >>> wn.synset('dog.n.01').lemmas
  - [Lemma('dog.n.01.dog'), Lemma('dog.n.01.domestic_dog'), Lemma('dog.n.01.Canis_familiaris')]
  - >>> [lemma.name for lemma in wn.synset('dog.n.01').lemmas]
  - ['dog', 'domestic_dog', 'Canis_familiaris']
  - >>> wn.lemma('dog.n.01.dog').synset

Wordnet in NLTK

- Hypernyms:
  - >>> dog1 = wn.synset('dog.n.01')
  - >>> dog1.hypernyms()
  - [Synset('domestic_animal.n.01'), Synset('canine.n.02')]
- Hyponyms:
  - >>> dog1.hyponyms()
  - [Synset('puppy.n.01'), Synset('great_pyrenees.n.01'), Synset('basenji.n.01'), ...]
- Hypernym closure: ...

Word Similarity

Outline

- Motivations
- Thesaurus-based measures
- Distributional measures
- Evaluation

Motivations

- Synonymy is a binary relation
  - Two words are either synonymous or not
- For many applications, we want a looser metric
  - Word similarity or word distance
- Informally: two words are more similar if they share more "features" of meaning
- Similarity and distance are relations between senses:
  - bank1 is similar to fund3 rather than "bank is like fund"
- We'll compute them over both words and senses

Why word similarity

- Information retrieval
- Question answering
- Machine translation
- Natural language generation
- Automatic essay grading
- Plagiarism detection
- Document classification/clustering
- ...

Two main classes of algorithms

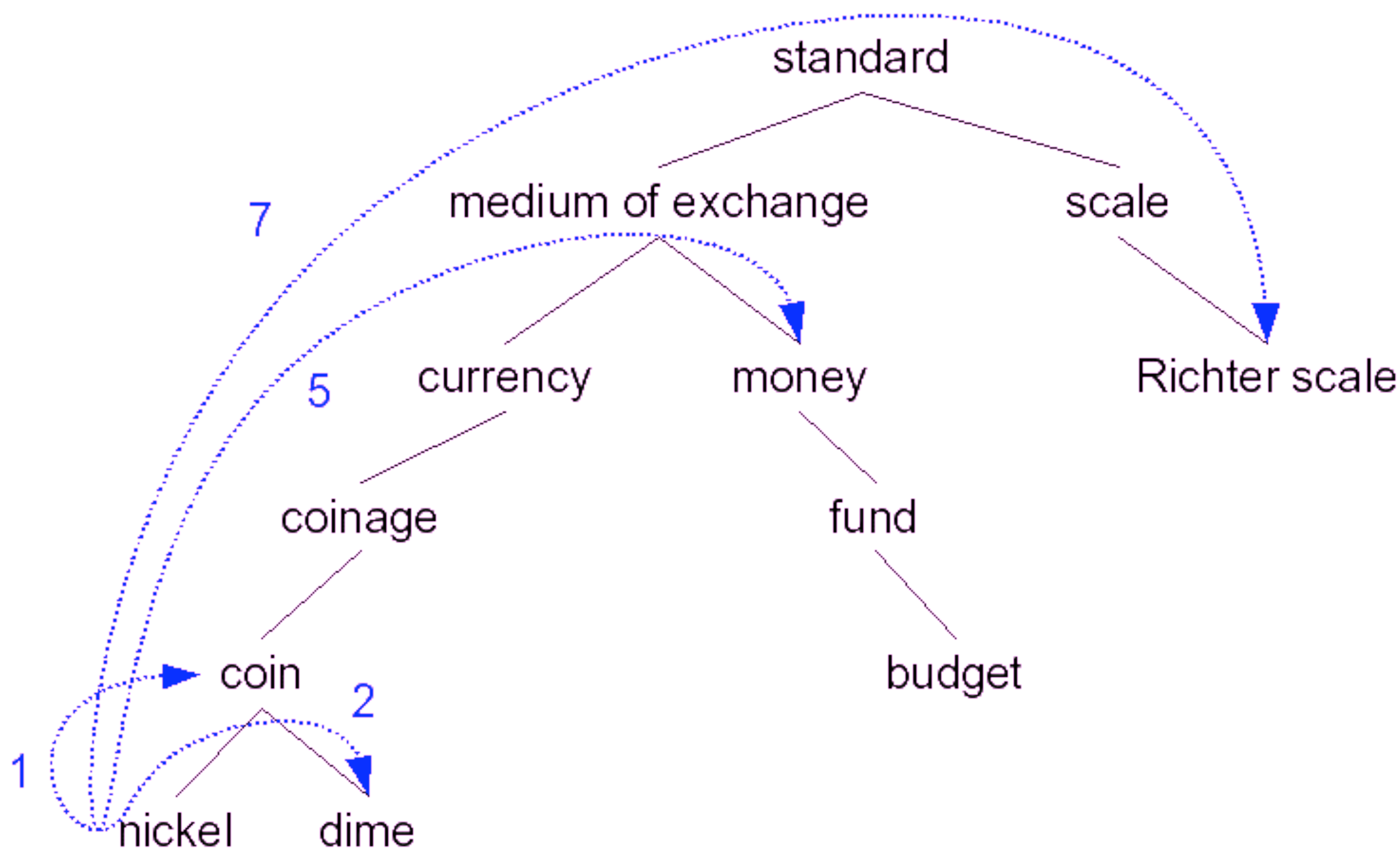
- Thesaurus-based algorithms
  - Words are compared in terms of how "close" they are in the thesaurus (e.g., Wordnet)
  - Requires a thesaurus (and a corpus)
- Distributional algorithms
  - Words are compared in terms of the shared number of contexts they can appear in
  - Requires (a lot of!) text but no thesaurus

Thesaurus-based word similarity

- We could use anything in the thesaurus:
  - Meronymy, glosses, example sentences
- In practice, we only use the is-a/subsumption/hypernym hierarchy
- Word similarity vs. word relatedness
  - Similar words are near-synonyms
  - Related words are in the same "semantic field"
    - Car, gasoline: related
    - Car, bicycle: similar

Path-based similarity

- Two words are similar if "nearby" in thesaurus hierarchy (i.e. short path between them)



Path-based similarity

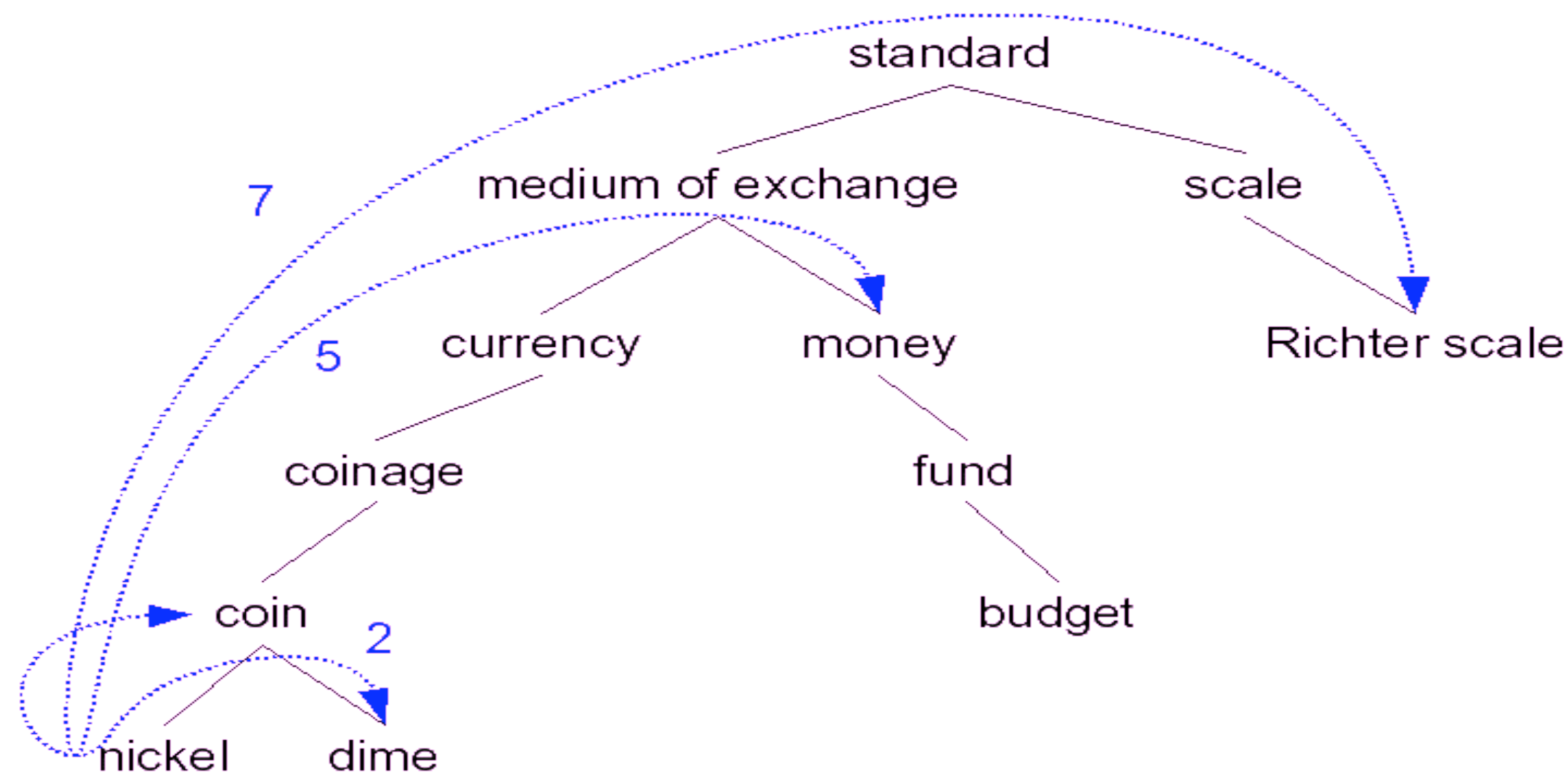
- Basic algorithm for path similarity:
  - compute # of edges in the shortest IS-A path in the thesaurus graph between the sense nodes c_1 and c_2
  - $sim_{path}(c_1, c_2) = -\log\ pathlength(c_1, c_2)$

- Variants of this measure proposed by: *Hirst and St-Onge (1998), Leacock & Chodorov (1998) ,Wu and Palmer (1994)*
- One can approximate word similarity by taking the most similar sense pair *(Resnik, 1995)*

$$wordsim(w_1, w_2) = \max_{c_1 \in senses(w_1), c_2 \in senses(w_2)} sim_{path}(c_1, c_2)$$

Problem with basic path-based similarity

- Assumes each link represents a uniform distance
- “nickel”-“money” seems closer than “nickel”-“standard”



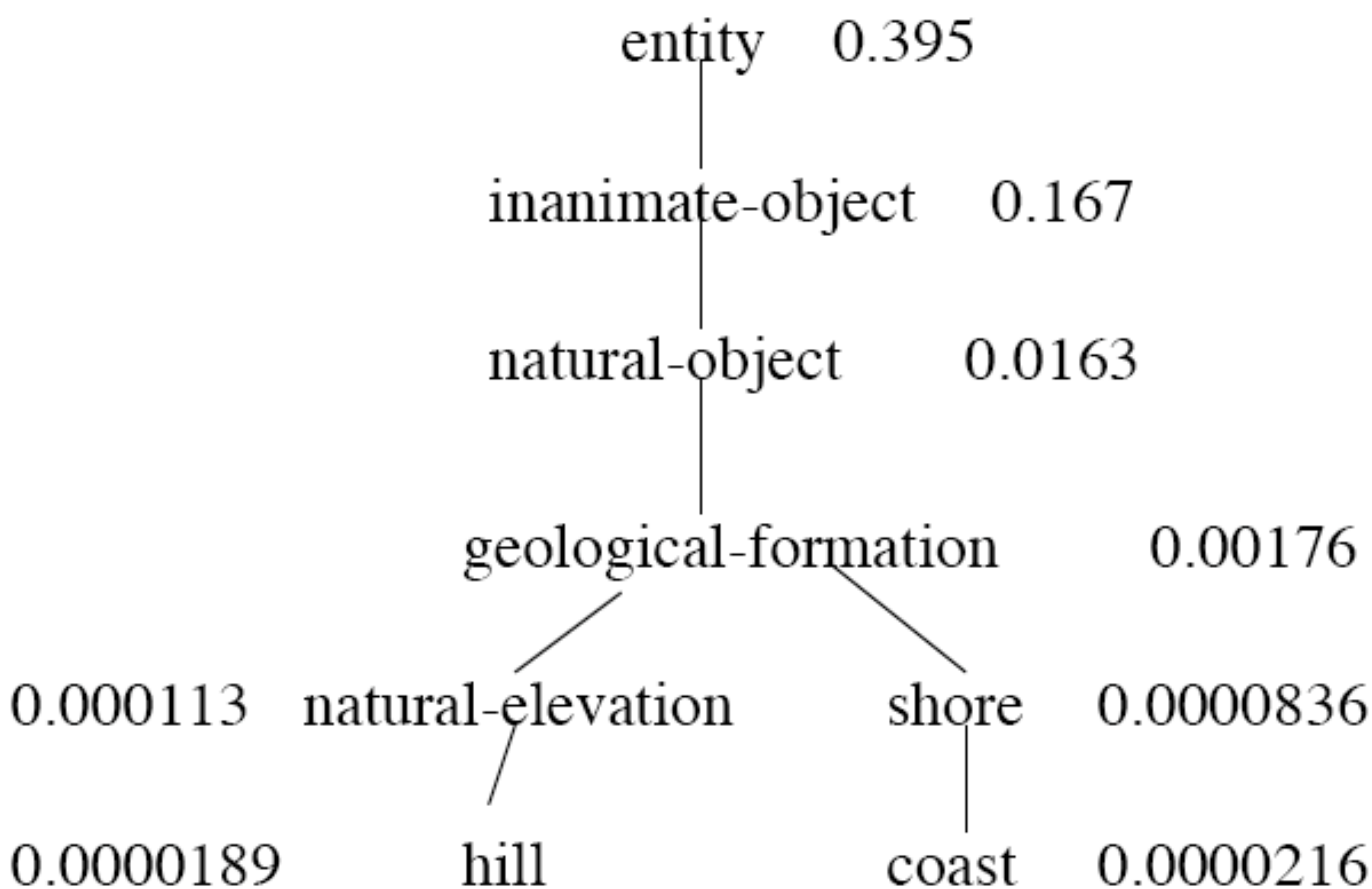
- Need a finer-grained metric which lets us represent the distance of each edge independently

Thesaurus-based similarity using corpus statistics

- Idea: use the structure of thesaurus and add probabilistic information derived from a corpus
  - Let's define $P(c)$ as:
    - The probability that a randomly selected word in a corpus is an instance of concept c
    - $$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$
  - The lower a node in the hierarchy, the lower its probability
  - A given word appearing in corpus counts toward frequency of all its hypernyms

Thesaurus-based similarity using corpus statistics

- Wordnet hierarchy augmented with probabilities $P(C)$



Information Content (IC) similarity

- Similarity between two words is related to the amount of information they have in common
- Information content:
  - $IC(c)$=-log $P(c)$
- Lowest common subsumer, $LCS(c1,c2)$:
  - the lowest node in the hierarchy
  - that subsumes (is a hypernym of) both $c1$ and $c2$
- Resnik (1995) 's similarity:
  - $sim_{resnik}(c_1, c_2) = -\log P(LCS(c_1, c_2))$

Information Content (IC) similarity measures

- *Resnik (1995)*'s similarity:

$$sim_{resnik}(c_1, c_2) = -\log P(LCS(c_1, c_2))$$

- *Lin (1998)*'s similarity:

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

- *Jiang and Conrath (1997)*'s similarity:

- $$sim_{JC}(c_1, c_2) = \frac{1}{2\log P(LCS((c_1, c_2)) - (\log P(c_1) + \log P(c_2))}$$

Dictionary-based similarity: Extended Lesk Algorithm

- Hypothesis: two concepts are similar if their glosses contain similar words
  - *drawing paper*: paper that is specially prepared for use in drafting
  - *decal*: the art of transferring designs from specially prepared paper to a wood or glass or metal surface
- For each *n*-word phrase that occurs in both glosses
  - Add a score of n^2 (to favor multi-word overlaps)
  - *paper* and *specially prepared* gives $1^2 + 2^2 = 5$…
- Extented Lesk also computes overlaps between hypernyms, hyponyms, meronyms glosses

Summary: thesaurus-based similarity

$$\begin{aligned}
\mathrm{sim}_{\mathrm{path}}(c_1,c_2) &= -\log \mathrm{pathlen}(c_1,c_2) \\
\mathrm{sim}_{\mathrm{Resnik}}(c_1,c_2) &= -\log P(\mathrm{LCS}(c_1,c_2)) \\
\mathrm{sim}_{\mathrm{Lin}}(c_1,c_2) &= \frac{2 \times \log P(LCS(c_1,c_2))}{\log P(c_1) + \log P(c_2)} \\
\mathrm{sim}_{\mathrm{jc}}(c_1,c_2) &= \frac{1}{2 \times \log P(\mathrm{LCS}(c_1,c_2)) - (\log P(c_1) + \log P(c_2))} \\
\mathrm{sim}_{\mathrm{eLesk}}(c_1,c_2) &= \sum_{r,q \in \mathrm{RELS}} \mathrm{overlap}(\mathrm{gloss}(r(c_1)), \mathrm{gloss}(q(c_2)))
\end{aligned}$$

- NLTK implements these metrics and other thesaurus-based metrics in its wordnet module.

Evaluating thesaurus-based similarity

- Intrinsic Evaluation:
  - Correlation coefficient between algorithm scores and word similarity ratings from humans
- Extrinsic (task-based, end-to-end) Evaluation:
  - Embed in some end application
    - Malapropism (spelling error) detection
    - Essay grading
    - Plagiarism Detection
    - Language modeling in some application
- Jiang-Conrath and Extended Lesk perform best

Problems with thesaurus-based methods

- We don't have a thesaurus for every language
- Even if we do, many words are missing (e.g., new words, domain specific words)
- Mostly, they rely on hyponym info:
  - Strong for nouns, but lacking for adjectives and even verbs
- Alternative
  - Distributional methods for word similarity

Distributional methods for word similarity

- Firth (1957): "You shall know a word by the company it keeps!"
- Nida (1975)'s example noted by Lin (1998):
  - A bottle of **tezgüino** is on the table
  - Everybody likes **tezgüino**
  - **Tezgüino** makes you drunk
  - We make **tezgüino** out of corn.
- Intuition:
  - just from these contexts a human could guess meaning of <u>tezguino</u>
  - So we should look at the surrounding contexts, see what other words have similar context.

Word meaning as context vector

- Consider a target word w
- Suppose we had one binary feature f_i for each of the N words in the lexicon v_i
  - Which means "word v_i occurs in the neighborhood of w"
- Word meaning represented as context vector:
  - $\mathbf{w} = (f1, f2, f3, \ldots, fN)$
- If *w=tezguino, v1=bottle, v2=drunk, v3=matrix*:
  - $\mathbf{w} = (1,1,0,\ldots)$

Intuition

- Define two words by these sparse features vectors
- Apply a vector distance metric
- Say that two words are similar if two vectors are similar

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

Distributional similarity

- Three main things to specify:
  - 1. What's the most adequate representation of context?
    - How to define co-occurrence terms? Simple word co-occurrences or more refined?
  - 2. How do measure the association with context?
    - How do we weight the co-occurrence terms: binary, frequency, mutual information?
  - 3. How do we define similarity between co-occurrences vectors
    - Which vector distance metric should we use: Euclidean/ Manhattan distance, cosine, Jaccard?

1. Defining co-occurrence vectors

- We could have windows
  - Bag-of-words
  - We generally remove stopwords
- But the vectors are still very sparse...
- So instead of using ALL the words in the neighborhood, how about just the words occurring in particular grammatical relations (Hindle, 1990)
  - For example, works like *tea*, *water*, *beer* are all frequent direct objects of the verb *drink*.
- Good news is: there are lot of dependency parsers out there that can give us relations: subject, DO, IO, modifiers, ...

I. Defining co-occurrence vectors

- Each dependency parse gives us a set of dependency tuples (= our contexts or features)

I discovered dried tangerines:

discover (subject I) I (subj-of discover)
tangerine (obj-of discover) tangerine (adj-mod dried)
dried (adj-mod-of tangerine)

- These tuples are used to build co-occurrence vectors:

	subj-of, absorb	subj-of, adapt	subj-of, behave	...	pobj-of, inside	pobj-of, into	...	nmod-of, abnormality	nmod-of, anemia	nmod-of, architecture	...	obj-of, attack	obj-of, call	obj-of, come from	obj-of, decorate	...	nmod, bacteria	nmod, body	nmod, bone marrow
cell	1	1	1		16	30		3	8	1		6	11	3	2		3	2	2

2. Weighting the counts

- We have our features/word's contexts, but we still don't know how to weight them
- Some options:
  - Binary values
    - $\mathrm{assoc}_{binary}(w,f)$ = 0 or 1 if word appears in context *f*
  - Frequency:
    - $\mathrm{assoc}_{prob}(w,f)$ = *P(f,w)* = count(w,f)/count(w') (where w' are all the words appearing in context f = (r,w'))
- Too coarse:
  - These schemes are not good at distinguishing informative contexts from uninformative ones: *(has-obj water)*/*(has-obj it)*
- We need a measure that asks how much more often than chance the feature co-occurs with the word

2. Weighting the counts: Mutual Information

- Mutual information: between 2 random variables X & Y

$$I(X,Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- Pointwise mutual information (PMI): measures how often two events x and y occur, compared with what we would expect if they were independent:

- $$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

2. Weighting the counts: Mutual Information

- PMI between a target word w and a feature f :

$$\text{assoc}_{\text{PMI}}(w,f) = \log_2 \frac{P(w,f)}{P(w)P(f)}$$

- Lin (1998) measure is a variant of PMI, breaks down expected value for P(f) differently:

$$\text{assoc}_{\text{Lin}}(w,f) = \log_2 \frac{P(w,f)}{P(w)P(r|w)P(w'|w)}$$

2. Weighting the counts: PMI rather than frequency

- "drink it" is more common than "drink wine"
- But "wine" is a better "drinkable" thing than "it"
- Idea:
  - We need to control for change (expected frequency)
  - We do this by normalizing by the expected frequency we would get assuming independence

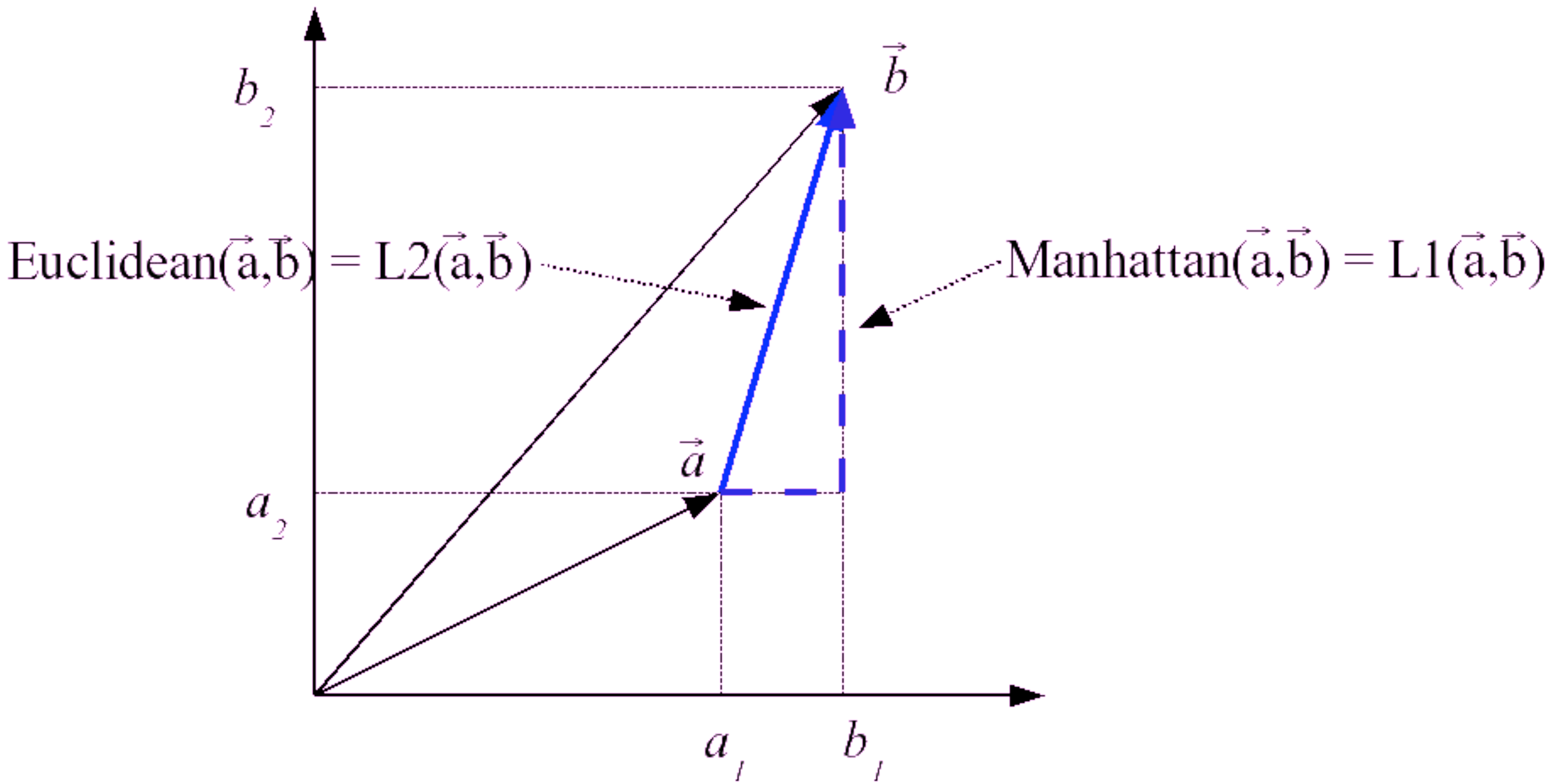
Object	Count	PMI assoc	Object	Count	PMI assoc
bunch beer	2	12.34	wine	2	9.34
tea	2	11.75	water	7	7.65
Pepsi	2	11.75	anything	3	5.15
champagne	4	11.75	much	3	5.15
liquid	2	10.53	it	3	1.25
beer	5	10.20	<SOME AMOUNT>	2	1.22

2. Weighting the counts: other measures

- See Manning and Schuetze (1999) for more

$$
\begin{aligned}
\mathrm{assoc}_{\mathrm{prob}}(w,f) &= P(f|w) \\
\mathrm{assoc}_{\mathrm{PMI}}(w,f) &= \log_2 \frac{P(w,f)}{P(w)P(f)} \\
\mathrm{assoc}_{\mathrm{Lin}}(w,f) &= \log_2 \frac{P(w,f)}{P(w)P(r|w)P(w'|w)} \\
\mathrm{assoc}_{\text{t-test}}(w,f) &= \frac{P(w,f)-P(w)P(f)}{\sqrt{P(f)P(w)}}
\end{aligned}
$$

3. Similarity between vectors



3. Similarity between vectors

$$\mathrm{sim}_{\mathrm{cosine}}(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

$$\mathrm{sim}_{\mathrm{Jaccard}}(\vec{v},\vec{w}) = \frac{\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} \max(v_i,w_i)}$$

$$\mathrm{sim}_{\mathrm{Dice}}(\vec{v},\vec{w}) = \frac{2\times\sum_{i=1}^{N} \min(v_i,w_i)}{\sum_{i=1}^{N} (v_i+w_i)}$$

$$\mathrm{sim}_{\mathrm{JS}}(\vec{v}||\vec{w}) = D(\vec{v}|\frac{\vec{v}+\vec{w}}{2}) + D(\vec{w}|\frac{\vec{v}+\vec{w}}{2})$$

Evaluating similarity

- Intrinsic Evaluation:
  - Correlation coefficient between algorithm scores and word similarity ratings from humans
- Extrinsic (task-based, end-to-end) Evaluation:
  - Malapropism (spelling error) detection
  - WSD
  - Essay grading
  - Taking TOEFL multiple-choice vocabulary tests
  - Language modeling in some application