Information Retrieval (IN4325)

Semantics

Dr. Nava Tintarev

Assistant Professor, TU Delft

Credits: Many of these slides are modified from the Stanford NLP course:
https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html

TU Delft

Last week...

Probabilistic language models

n-grams

Smoothing

Interpolation

(Stupid) back-off

Using syntax in sentiment analysis

TU Delft

This week

Semantics

- Word senses
- Path based similarity
- Information content similarity
- Distributional similarity

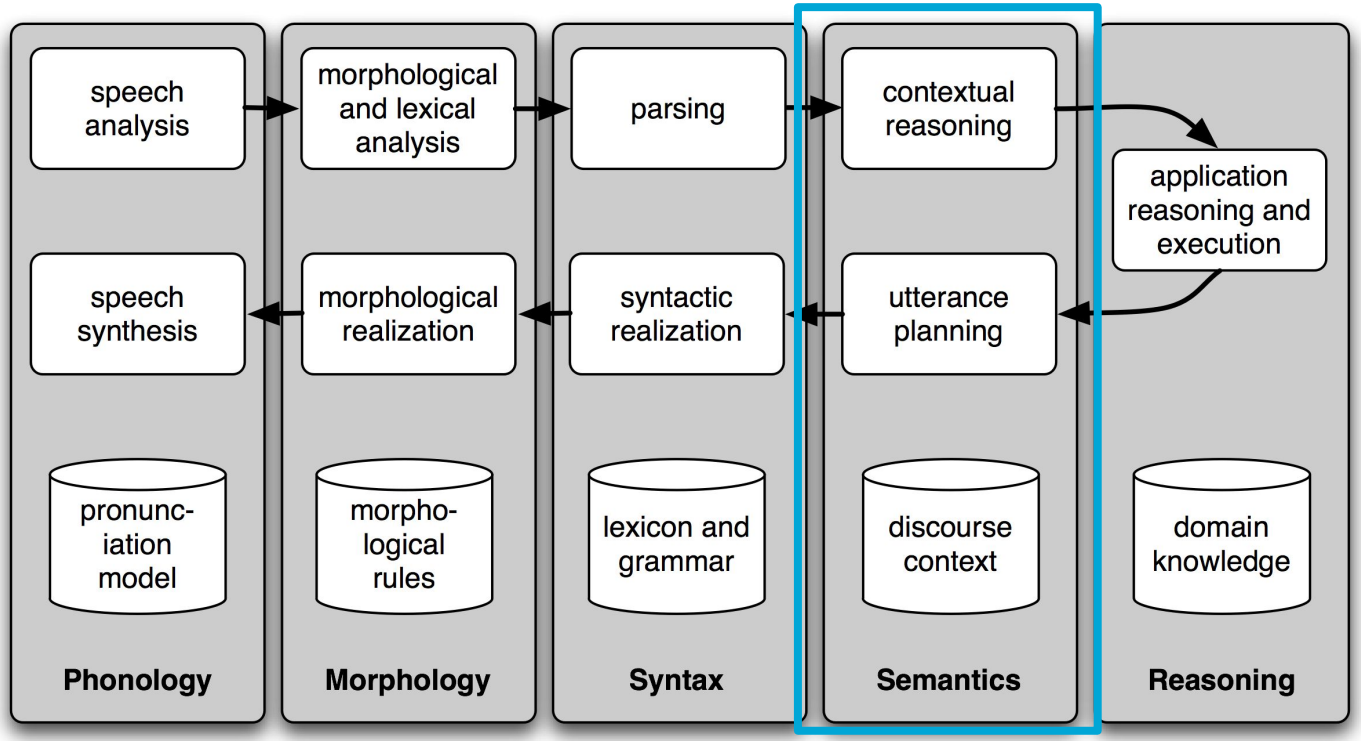
Applications

- Sentiment analysis
- Question answering
- NLG - Lexical choice

TU Delft

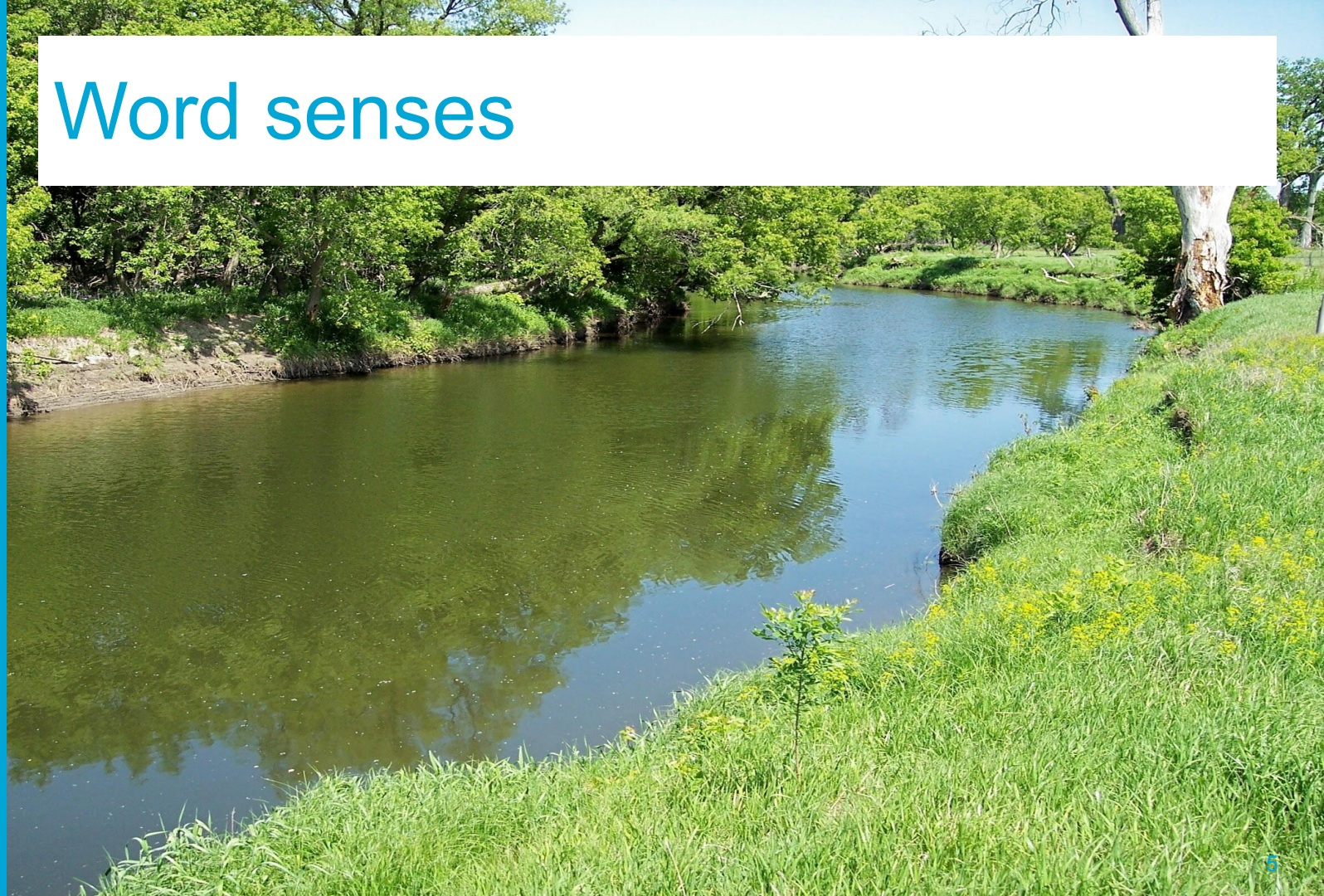
Sub-tasks

What does it all mean?



Representation of meaning

Word senses



TU Delft

Reminder: lemma and wordform

- A **lemma** or **citation form**
  - Same stem, part of speech, rough semantics
- A **wordform**
  - The “inflected” word as it appears in text

Wordform	Lemma
banks	bank
sung	sing
duermes	dormir

Lemmas have senses

- One lemma “bank” can have many meanings:
  - Sense 1: …a **bank**[1] can hold the investments in a custodial account…
  - Sense 2: “…as agriculture burgeons on the east **bank**[2] the river will shrink even more”
- **Sense** (or **word sense**)
  - A discrete representation of an aspect of a word’s meaning.
- The lemma **bank** here has two senses

Name

Writing

Sound

Homonymy

Homonyms: words that share a form but have unrelated, distinct meanings:

- $bank_1$: financial institution, $bank_2$: sloping land
- bat_1: club for hitting a ball, bat_2: nocturnal flying mammal

Homographs: "It's not **use** to ask to **use** the telephone"; "Do you **live** near zoo with **live** animals"

Homophones:

1. Write and right
2. Piece and peace

TU Delft

Homonymy causes problems for NLP applications

- Information retrieval
  - "bat care"
- Machine Translation
  - bat: murciélago (animal) or bate (for baseball)
- Text-to-Speech
  - bass (stringed instrument) vs. bass (fish)

TU Delft

Polysemy

1. The **bank** was constructed in 1875 out of local red brick.
2. I withdrew the money from the **bank**

Are those the same sense?

- Sense 1: “The building belonging to a financial institution”
- Sense 2: “A financial institution”

Like polymorphism. ..

- A **polysemous** word has **related** meanings
  - Most non-rare words have multiple meanings

TU Delft

Metonymy or Systematic Polysemy: A systematic relationship between senses

- Lots of types of polysemy are systematic
  - `School, university, hospital`
  - All can mean the institution or the building.
- A systematic relationship:
  - Building ⟷ Organization
- Other such kinds of systematic polysemy:

Author (`Jane Austen wrote Emma`)
⟷ Works of Author (`I love Jane Austen`)

Tree (`Plums have beautiful blossoms`)
⟷ Fruit (`I ate a preserved plum`)

How do we know when a word has more than one sense?

ζεῦγμα, *zeûgma*, lit. "a yoking together"

- The “zeugma” test: Two senses of `serve`?
  - `Which flights` **`serve`** `breakfast?`
  - `Does Lufthansa` **`serve`** `Amsterdam?`
  - **?Does Lufthansa serve breakfast and Amsterdam?**

- Since this conjunction sounds weird,
  - we say that these are **two different senses of “serve”**

TU Delft

Similarity

Synonyms

- **Word that have the same meaning in some or all contexts.**
  - couch / sofa
  - big / large
  - automobile / car
  - vomit / throw up
  - Water / H_20

- **Two lexemes are synonyms**
  - if they can be substituted for each other in all situations
  - If so they have the same **propositional meaning**

Synonyms

- NB: There are few (or no) examples of perfect synonymy.
  - Even if many aspects of meaning are identical
  - Still may not preserve the acceptability based on notions of politeness, slang, register, genre, etc.

Synonymy is a 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 became a kind of **big** sister to Benjamin.
  - ?Miss Nelson became a kind of **large** sister to Benjamin.
- Why?
  - *big* has a sense that means being older, or grown up
  - *large* lacks this sense

Antonyms

- Senses that are opposites with respect to one feature of meaning
- Otherwise, they are very similar!

  `dark/light  short/long  fast/slow rise/fall`
  `hot/cold  up/down`

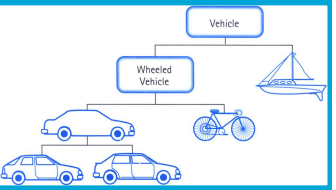
- More formally: **antonyms** can
  - define a binary opposition or be at opposite ends of a scale
    - `long/short, fast/slow`
  - Be **reversives** (describe a change in direction or movement):
    - `rise/fall, up/down`

Hyponymy and Hypernymy

- 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*
  - *mango* is a hyponym of *fruit*
- Conversely **hypernym/superordinate** ("hyper is super")
  - *vehicle* is a **hypernym** of *car*
  - *fruit* is a hypernym of *mango*

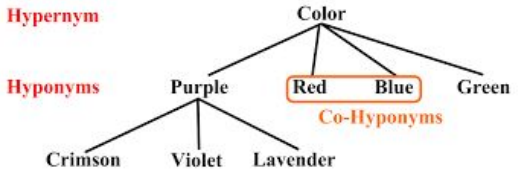
Superordinate/hyper	vehicle	fruit	furniture
Subordinate/hyponym	car	mango	chair

Similar to inheritance in OOP!



Hyponymy more formally

- Extensional:
  - The class **(color)** denoted by the superordinate/hypernym extensionally includes the class denoted by the hyponym (red)
- 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; B hypo C **--entails-->** A hypo C)
- Another name: the **IS-A hierarchy**
  - A IS-A B (or A ISA B)
  - B **subsumes** A



Hyponyms and Instances

- WordNet has both **classes** and **instances**.
- An **instance** is an individual, a proper noun that is a unique entity
    - `San Francisco` is an **instance** of `city`
  - But `city` is a class
    - `city` is a **hyponym** of `municipality...location...`

Applications of Thesauri and Ontologies

- Information Extraction
- Question Answering
- Bioinformatics and Medical Informatics
- Machine Translation

WordNet 3.0

- A hierarchically organized lexical database
- On-line thesaurus + aspects of a dictionary
    - Some other languages available or under development

DutchSemCor project

- 250K tagged manually, 500 million words (WSD systems)
- http://wordpress.let.vupr.nl/dutchsemcor/

Category	Unique Strings
Noun	117,798
Verb	11,529
Adjective	22,479
Adverb	4,481

TU Delft

Senses of “bass” in Wordnet

Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- **S: (n) bass, basso (an adult male singer with the lowest voice)**
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

How is “sense” defined in WordNet?

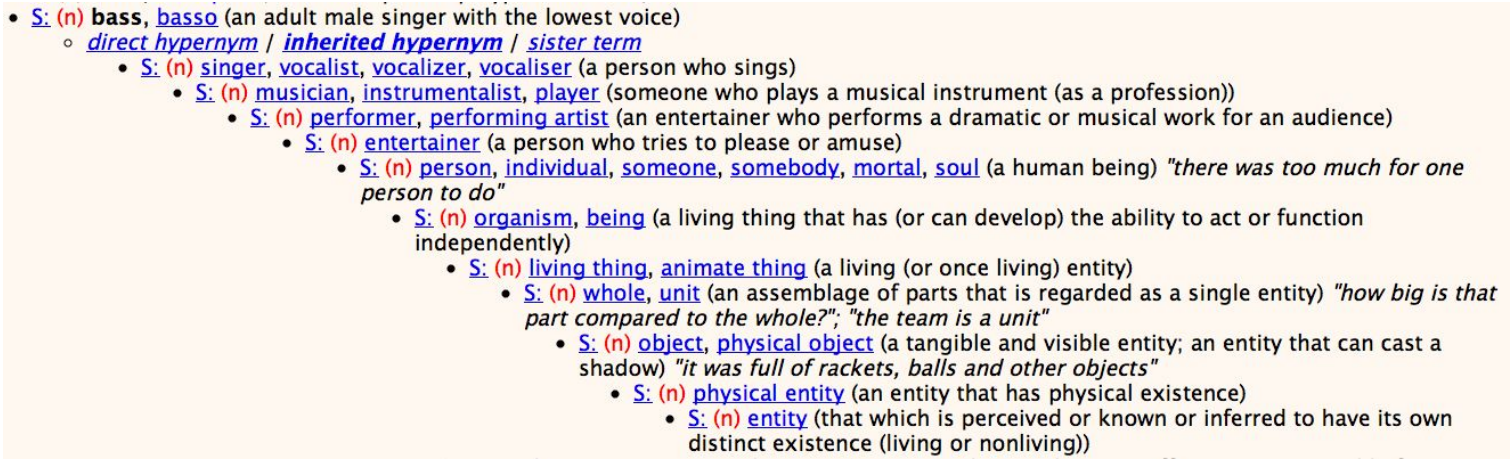
- The **synset (synonym set)**, the set of near-synonyms, instantiates a sense or concept, with a gloss.
- **Example:** chump as a noun with the gloss:

  `“a person who is gullible and easy to take advantage of”`
- This sense of “chump” is shared by 9 words:

  `chump`[1], `fool`[2], `gull`[1], `mark`[9], `patsy`[1], `fall guy`[1], `sucker`[1], `soft touch`[1], `mug`[2]
- Each of **these** senses have this same gloss
  - (Not **every** sense; sense 2 of gull is the aquatic bird)

TU Delft

WordNet Hypernym Hierarchy for “bass”



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

- **Where it is:**
  - http://wordnetweb.princeton.edu/perl/webwn
- **Libraries**
  - Python: WordNet from NLTK
    - http://www.nltk.org/Home
  - Java:
    - JWNL, extJWNL on sourceforge
- More on the Wordnet page….

Word Similarity

- **Synonymy**: a binary relation
  - Two words are either synonymous or not
- **Similarity** (or **distance**): a looser metric
  - Two words are more similar if they share more features of meaning
- Similarity is properly a relation between **senses**
  - The word "`bank`" is not similar to the word "`slope`"
  - Bank[1] is similar to fund[3]
  - Bank[2] is similar to slope[5]
- But we'll compute similarity over both words and senses

Why word similarity

- Question answering
- Machine translation
- Natural language generation
- Language modeling
- Automatic essay grading
- Plagiarism detection
- Document clustering

Word similarity and word relatedness

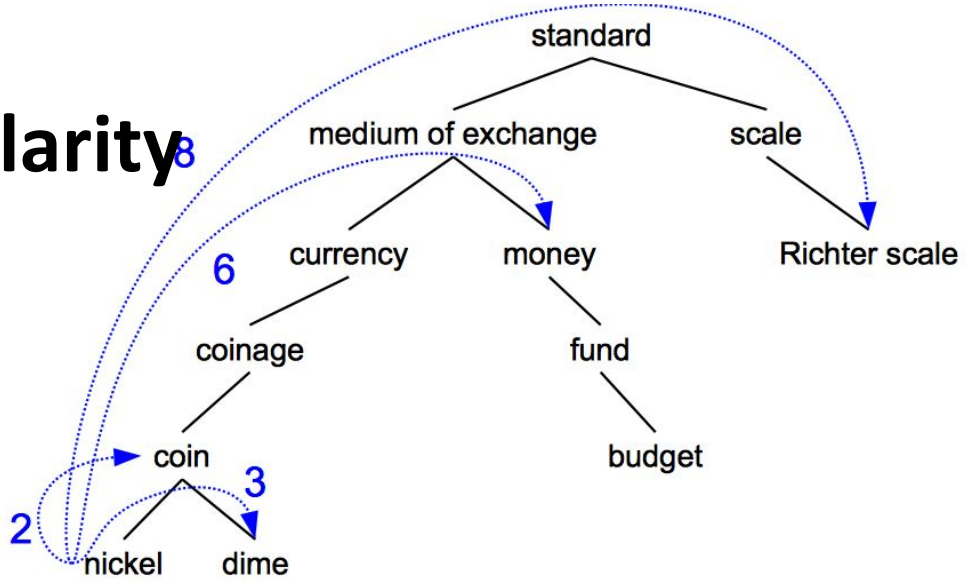
- We often distinguish **word similarity** from **word relatedness**
  - **Similar words**
    - `car, bicycle:` **similar**
  - **Related words**: can be related any way
    - `car, wheel, gasoline:` **related**, not similar

Two classes of similarity algorithms

- Thesaurus-based algorithms
  - Are words “nearby” in hypernym hierarchy?
  - Do words have similar glosses (definitions)?
- Distributional algorithms
  - Do words have similar distributional contexts?

Path-based similarity

Path based similarity



- Two concepts (senses/synsets) are similar if they are near each other in the thesaurus hierarchy
  - =have a short path between them
  - concepts have path 1 to themselves

Refinements to path-based similarity

- **pathlen(c_1,c_2)** = 1 + number of edges in the shortest path in the hypernym graph between sense nodes c_1 and c_2
- **simpath(c_1,c_2)** = $$\frac{1}{\text{pathlen}(c_1, c_2)}$$

ranges from 0 to 1 (identity)

- **wordsim(w_1,w_2)** = $$\max_{c_1 \in \text{senses}(w_1), c_2 \in \text{senses}(w_2)} \text{sim}(c_1, c_2)$$

Example: path-based similarity

$$\text{simpath}(c_1, c_2) = 1/\text{pathlen}(c_1, c_2)$$

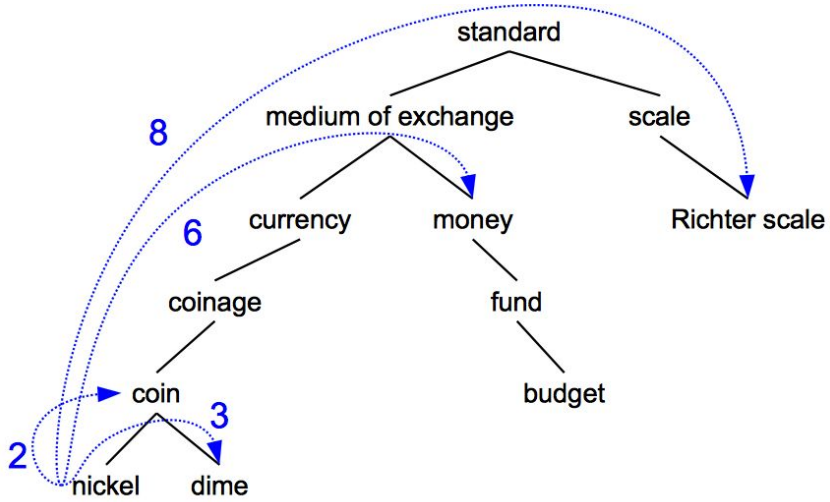
simpath(*nickel,coin*) = 1/2 =.5

simpath(*fund,budget*) = 1/2 =.5

simpath(*nickel,currency*) = 1/4 =.25

simpath(*nickel,money*) = 1/6 = .17

simpath(*coinage,Richter scale*)
=1/6 =.17



Problem with basic path-based similarity

- Assumes each link represents a uniform distance
  - But *nickel* to *money* seems to us to be closer than *nickel* to *standard*
  - Nodes high in the hierarchy are very abstract
- We instead want a metric that
  - Represents the cost of each edge independently
  - Words connected only through abstract nodes are less similar

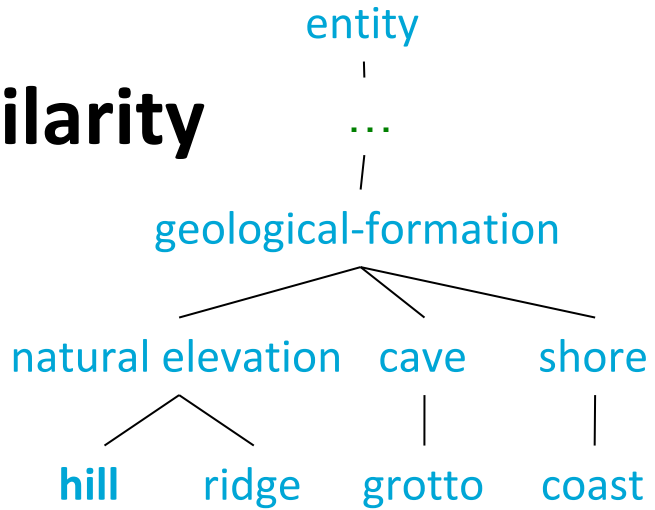
Information content similarity metrics

Resnik 1995. Using information content to evaluate semantic similarity in a taxonomy. IJCAI

- Let's define $P(c)$ as:
  - The probability that a randomly selected word in a corpus is an instance of concept *c*
  - Formally: there is a distinct random variable, ranging over words, associated with each concept in the hierarchy
    - for a given concept, each observed noun is either
      - **a member of that concept** with probability P(c)
      - **not** a member of that concept with probability 1-P(c)
  - All words are members of the root node (Entity)
    - P(root)=1
  - The lower a node in hierarchy, the lower its probability

Information content similarity

- Train by counting in a corpus
  - Each instance of ***hill*** counts toward frequency of *natural elevation, geological formation, entity*, etc
  - Let $\text{words}(c)$ be the set of all words that are children of node c
    - words("geological-formation") = {hill,ridge,grotto,coast,cave,shore, natural elevation}
    - words("natural elevation") = {hill, ridge}

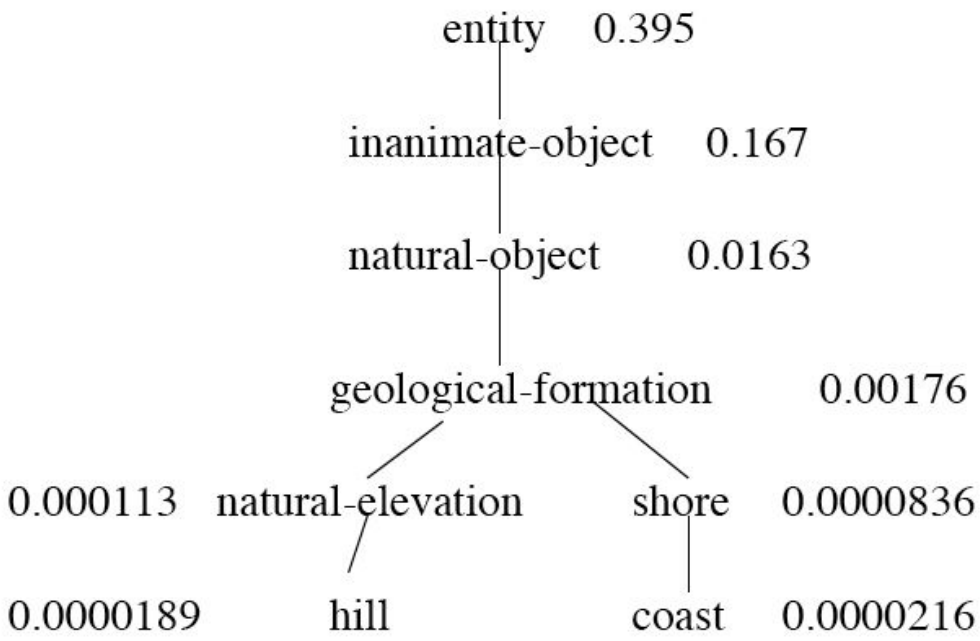


$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$

Information content similarity

- WordNet hierarchy augmented with probabilities P(c)

D. Lin. 1998. An Information-Theoretic Definition of Similarity. ICML 1998

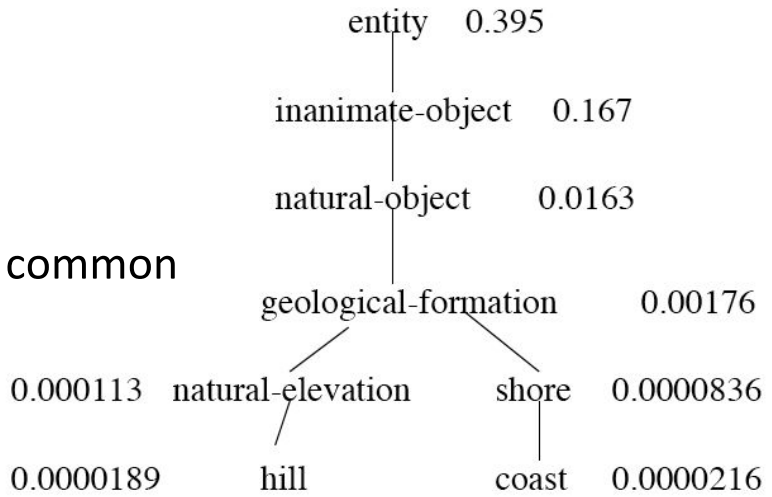


Information content: definitions

- Information content:

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

- Most informative subsumer (Lowest common subsumer)



$LCS(c_1,c_2) =$

The most informative (lowest) node in the hierarchy subsuming both $\mathbf{c_1}$ and $\mathbf{c_2}$.

Using information content for similarity: the Resnik method

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. IJCAI 1995.
Philip Resnik. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. JAIR 11, 95-130.

- The similarity between two words is related to their common information.
- The more two words have in common, the more similar they are.
- Resnik: measure common information as:
  - The information content of the most informative (lowest) common subsumer of the two nodes
  - $\mathbf{sim}_{\mathbf{resnik}}(\mathbf{c_1},\mathbf{c_2}) = -\log P(\ LCS(c_1,c_2)\ )$
- Similarity between a concept and itself varies between concepts

Dekang Lin method

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. ICML

- Intuition: Similarity between A and B is not just what they have in common
- The more **differences** between A and B, the less similar they are:
  - Commonality: the more A and B have in common, the more similar they are
  - Difference: the more differences between A and B, the less similar

Dekang Lin similarity theorem

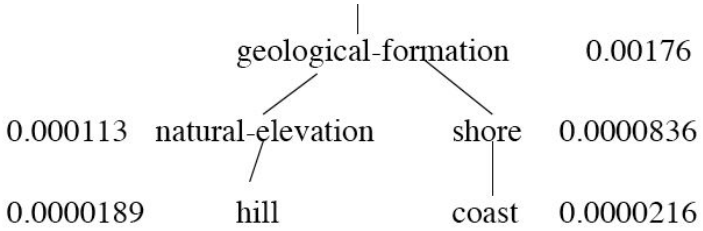
- The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are

$$sim_{Lin}(A,B) \propto \frac{IC(common(A,B))}{IC(description(A,B))}$$

- Lin (altering Resnik) defines IC(common(A,B)) as 2 x information of the LCS.

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

Lin similarity function



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

$$sim_{Lin}(\text{hill},\text{coast}) = \frac{2\log P(\text{geological-formation})}{\log P(\text{hill})+\log P(\text{coast})}$$

$$= \frac{2\ln 0.00176}{\ln 0.0000189 + \ln 0.0000216}$$

$$= .59$$

The (extended) Lesk Algorithm

- A thesaurus-based measure that looks at **glosses**
- 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 is in both glosses
  - Add a score of n^2
  - `Paper` and `specially prepared`: $1 + 2^2 = 5$
  - Compute overlap also for other relations
    - glosses of hypernyms and hyponyms

Summary: thesaurus-based similarity

$$\text{sim}_{\text{path}}(c_1, c_2) = \frac{1}{pathlen(c_1, c_2)}$$

$$\text{sim}_{\text{resnik}}(c_1, c_2) = -\log P(LCS(c_1, c_2)) \quad \text{sim}_{\text{lin}}(c_1, c_2) = \frac{2\log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)}$$

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

$$\text{sim}_{eLesk}(c_1, c_2) = \sum_{r,q \in RELS} \text{overlap}(\text{gloss}(r(c_1)), \text{gloss}(q(c_2)))$$

Questions?

Libraries for computing thesaurus-based similarity

- NLTK
  - http://nltk.github.com/api/nltk.corpus.reader.html?highlight=similarity - nltk.corpus.reader.WordNetCorpusReader.res_similarity

- WordNet::Similarity
  - http://wn-similarity.sourceforge.net/
  - Web-based interface:
    - http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi

48

Evaluating similarity

- Intrinsic Evaluation:
  - Correlation between algorithm and human word similarity ratings
- Extrinsic (task-based, end-to-end) Evaluation:
  - Malapropism (spelling error) detection
  - Word sense disambiguation (WSD)
  - Essay grading
  - Taking TOEFL multiple-choice vocabulary tests

Levied is closest in meaning to:

imposed, believed, requested, correlated

Similar news titles?

Sean Connery fit after tumour operation.

a) James Bond on the mend.
b) Crowe's son names his brother.
c) NASA relieved as probe makes orbit.

Human judgments versus automatic judgements

- http://navatintarev.com/papers/TintarevMasthoffAH2006.pdf

Problems with thesaurus-based meaning

- We do not have a thesaurus for every language
- Even if we do, they have problems with **recall**
  - Many words are missing
  - Most (if not all) phrases are missing
  - Some connections between senses are missing
  - Thesauri work less well for verbs, adjectives
    - Adjectives and verbs have less structured hyponymy relations

Distributional models of meaning

- Also called vector-space models of meaning
- Offer much higher recall than hand-built thesauri
  - Although they tend to have lower precision
- Zellig Harris (1954): "**oculist** and **eye-doctor** … occur in almost the same environments…." **If A and B have almost identical environments we say that they are synonyms.**

- John Rupert Firth (1957): "*You shall know a word by the company it keeps*!"

52

Intuition of distributional word similarity

- Example:

> A bottle of ***tesgüino*** is on the table
> Everybody likes ***tesgüino***
> ***Tesgüino*** makes you drunk
> We make ***tesgüino*** out of corn.

- From context words humans can guess ***tesgüino*** means
  - an alcoholic beverage like **beer**
- Intuition for algorithm:
  - Two words are similar if they have similar word contexts.

The Term-Context matrix

- Instead of using entire documents, use smaller contexts
  - Paragraph
  - Window of 10 words
- A word is now defined by a vector over counts of context words

54

Should we use raw counts?

- For the term-document matrix
  - We use tf-idf instead of raw term counts
- For the term-context matrix
  - Positive Pointwise Mutual Information (PPMI) is common
  - **We will come back to this when we talk about word embeddings and Word2Vec….**

55

Pointwise Mutual Information

- **Pointwise mutual information**:

Do events x and y co-occur more than if they were independent?

$$\text{PMI}(X,Y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- **PMI between two words**: (Church & Hanks 1989)

Do words x and y co-occur more than if they were independent?

$$\text{PMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

- **Positive PMI between two words** (Niwa & Nitta 1994)
  - Replace all PMI values less than 0 with zero

$$ppmi_{ij} = \begin{cases} pmi_{ij} & \text{if } pmi_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Weighing PMI

- PMI is biased toward infrequent events
- Various weighting schemes help alleviate this
  - See Turney and Pantel (2010)
- Add-one smoothing can also help

57

Using syntax to define a word's context

- Zellig Harris (1968)
  - *"The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities"*
- Two words are similar if they have similar parse contexts
- **Duty** and **responsibility** (Chris Callison-Burch's example)

Modified by adjectives	additional, administrative, assumed, collective, congressional, constitutional …
Objects of verbs	assert, assign, assume, attend to, avoid, become, breach …

Co-occurrence vectors based on syntactic dependencies

Dekang Lin, 1998 “Automatic Retrieval and Clustering of Similar Words”

- The contexts C are different dependency relations
  - Subject-of- “absorb”
  - Prepositional-object of “inside”
- Counts for the word cell:

	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

PMI applied to dependency relations

Hindle, Don. 1990. Noun Classification from Predicate-Argument Structure. ACL

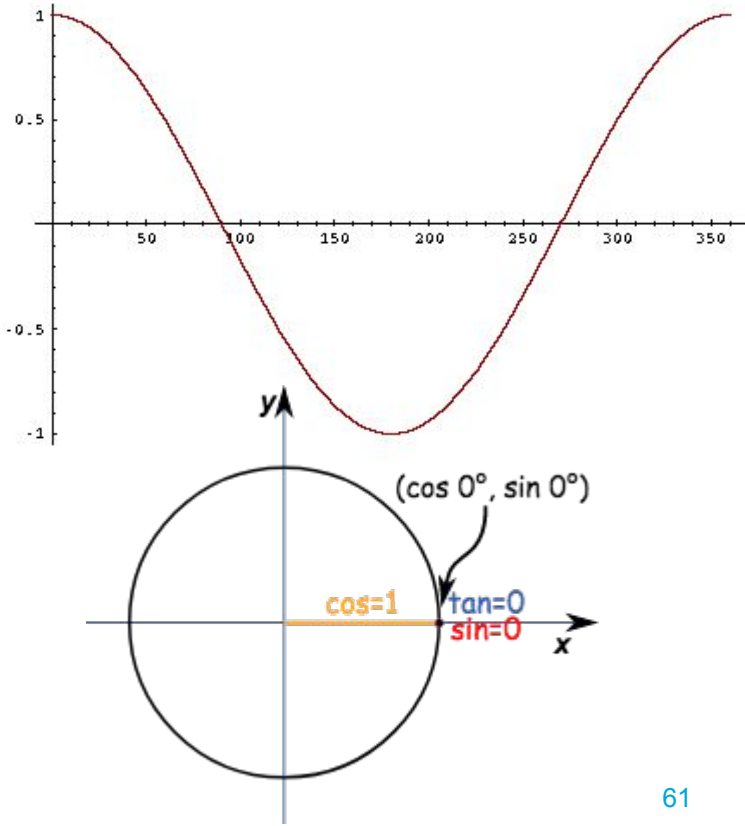
Object of "drink"	Count	PMI
tea	2	11.8
liquid	2	10.5
wine	2	**9.3**
anything	3	5.2
it	**3**	1.3

- "Drink it" more common than "drink wine" (Count)
- But "wine" is a better "drinkable" thing than "it" (PMI)

Cosine as a similarity metric

- -1: vectors point in opposite directions
- +1: vectors point in same directions
- 0: vectors are orthogonal

- Raw frequency or PPMI are non-negative, so cosine range 0-1



61

Other possible similarity measures

Jaccard - binary features, counts overlap.

Dice - Also binary. Denominator normalizes for number of non-zero entries

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

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

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

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

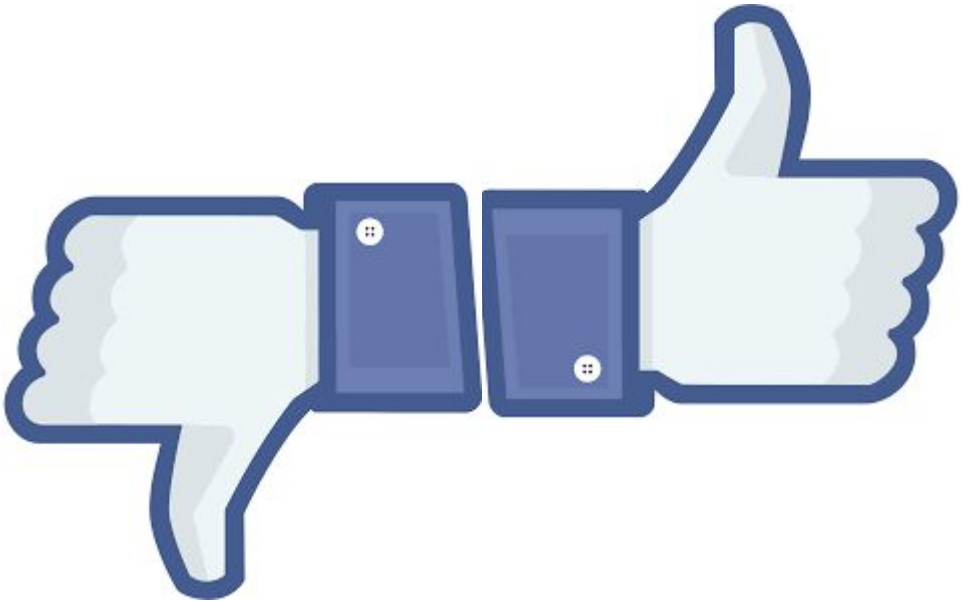
Jenson-Shannon - divergence of distributions from the mean of the two.

Evaluating similarity
(the same as for thesaurus-based)

- Intrinsic Evaluation:
  - Correlation between algorithm and human word similarity ratings
- Extrinsic (task-based, end-to-end) Evaluation:
  - Spelling error detection, WSD, essay grading
  - Taking TOEFL multiple-choice vocabulary tests

Levied is closest in meaning to which of these:
imposed, believed, requested, correlated

Applications



Sentiment: Using WordNet

S.M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. COLING 2004
M. Hu and B. Liu. Mining and summarizing customer reviews. In Proceedings of KDD, 2004

- Create positive (“good”) and negative seed-words (“terrible”)
- Find Synonyms and Antonyms
  - Positive Set: Add synonyms of positive words (“well”) and antonyms of negative words
  - Negative Set: Add synonyms of negative words (“awful”) and antonyms of positive words (”evil”)
- Repeat, following chains of synonyms
- Filter

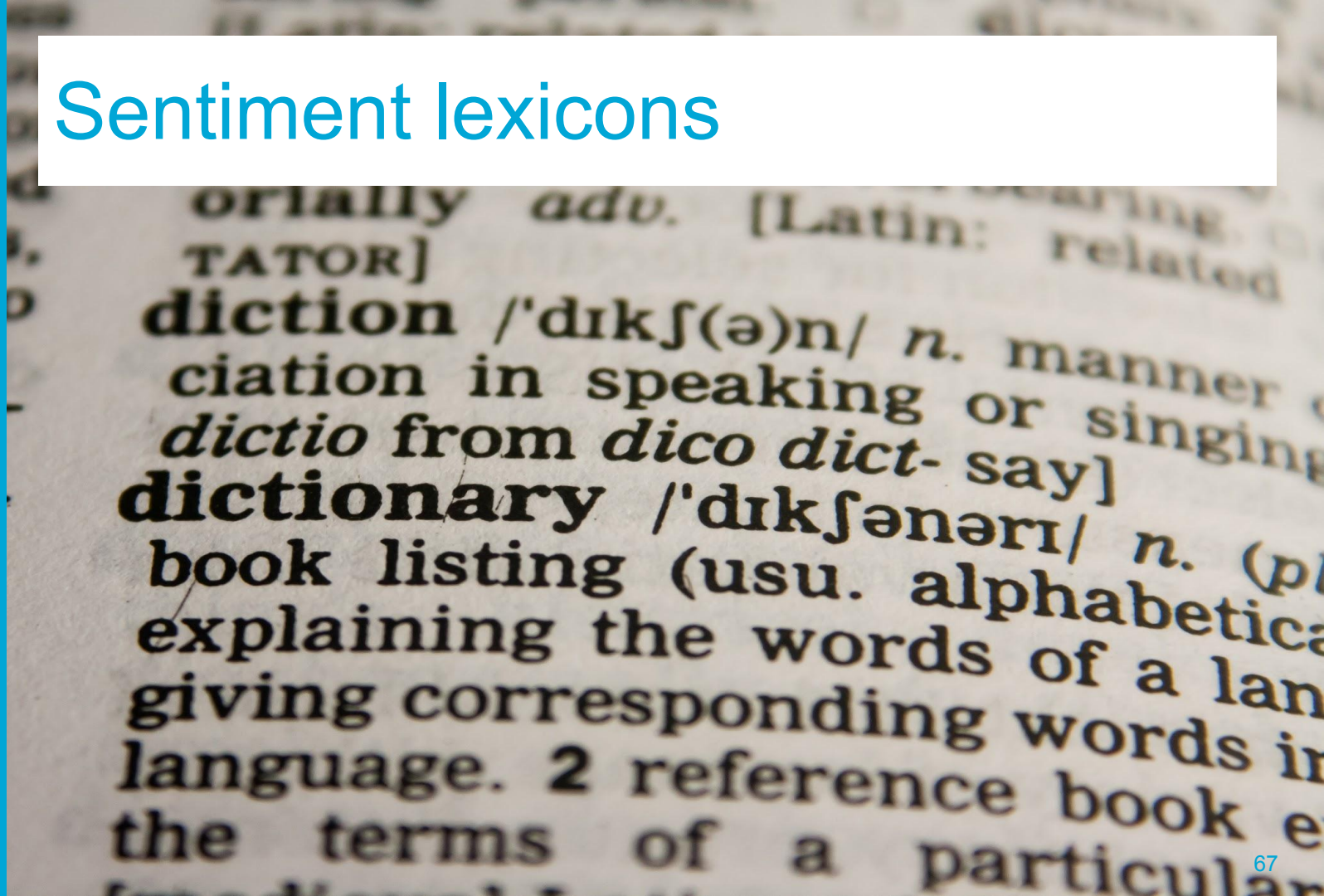
65

Problems:
What makes reviews hard to classify?

- Subtlety:
  - Perfume review in *Perfumes: the Guide*:
    - "If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut."
  - Dorothy Parker on Katherine Hepburn
    - "She runs the gamut of emotions from A to B"

66

Sentiment lexicons



The General Inquirer

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press

- Home page: http://www.wjh.harvard.edu/~inquirer
- List of Categories: http://www.wjh.harvard.edu/~inquirer/homecat.htm
- Spreadsheet: http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls
- Categories:
  - Positiv (1915 words) and Negativ (2291 words)
  - Strong vs Weak, Active vs Passive, Overstated versus Understated
  - Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc
- Free for Research Use

LIWC (Linguistic Inquiry and Word Count)

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

- Home page: http://www.liwc.net/
- 2300 words, >70 classes
- **Affective Processes**
  - negative emotion (*bad, weird, hate, problem, tough*)
  - positive emotion (*love, nice, sweet*)
- **Cognitive Processes**
  - Tentative (*maybe, perhaps, guess*), Inhibition (*block, constraint*)
- **Pronouns, Negation** (*no, never*), **Quantifiers** (*few, many*)
- $30 or $90 fee

TU Delft

MPQA Subjectivity Cues Lexicon

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). Recognizing Contextual Polarity in
Phrase-Level Sentiment Analysis. Proc. of HLT-EMNLP-2005.

Riloff and Wiebe (2003). Learning extraction patterns for subjective expressions. EMNLP-2003.

- Home page: http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- 6885 words from 8221 lemmas
  - 2718 positive
  - 4912 negative
- Each word annotated for intensity (strong, weak)
- GNU GPL

70

Bing Liu Opinion Lexicon

Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. ACM SIGKDD-2004.

- Bing Liu's Page on Opinion Mining
  - https://www.cs.uic.edu/~liub/
  - Opinion Lexicon (positive, negative)
  - Comparative lexicon

71

SentiWordNet

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010 SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-2010

- Home page: http://sentiwordnet.isti.cnr.it/
- All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness
- [estimable(J,3)] "may be computed or estimated"

  `Pos  0  Neg 0  Obj 1`
- [estimable(J,1)] "deserving of respect or high regard"

  `Pos .75  Neg 0  Obj .25`

Disagreements between polarity lexicons

Christopher Potts, Sentiment Tutorial, 2011

	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 **(0.6%)**	49/2867 **(2%)**	1127/4214 **(27%)**	12/363 **(3%)**
Opinion Lexicon		32/2411 **(1%)**	1004/3994 **(25%)**	9/403 **(2%)**
General Inquirer			520/2306 **(23%)**	1/204 **(0.5%)**
SentiWordNet				174/694 **(25%)**
LIWC				

Finding sentiment of a sentence

- Important for finding aspects or attributes
  - Target of sentiment

- `The food was great but the service was awful`

74

Finding aspect/attribute/target of sentiment

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In Proceedings of KDD.

S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. 2008. Building a Sentiment Summarizer for Local Service Reviews. WWW Workshop.

- Frequent phrases + rules
  - Find all highly frequent phrases across reviews ("`fish tacos`")
  - Filter by rules like "occurs right after sentiment word"
    - "...`great fish tacos`" means `fish tacos` a likely aspect

Casino	casino, buffet, pool, resort, beds
Children's Barber	haircut, job, experience, kids
Greek Restaurant	food, wine, service, appetizer, lamb
Department Store	selection, department, sales, shop, clothing

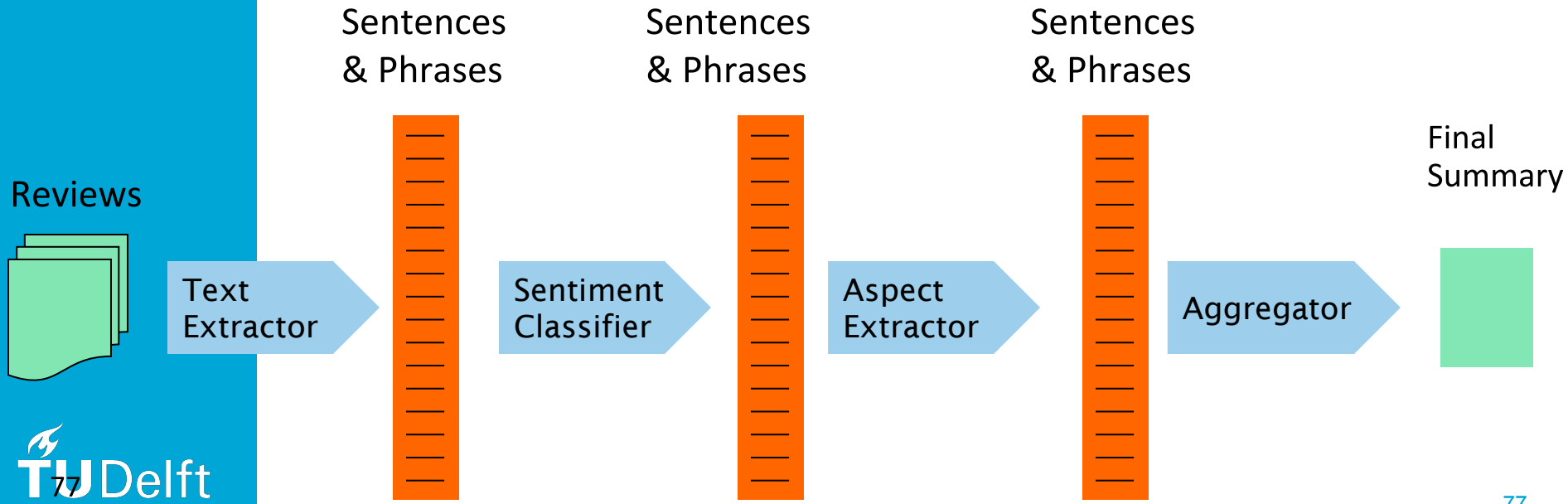
Finding aspect/attribute/target of sentiment

- The aspect name may not be in the sentence
- For restaurants/hotels, aspects are well-understood
- Supervised classification
  - Hand-label a small corpus of restaurant review sentences with aspect
    - food, décor, service, value, NONE
  - Train a classifier to assign an aspect to a sentence
    - "Given this sentence, is the aspect *food, décor, service, value,* or *NONE*"

76

Putting it all together: Finding sentiment for aspects

S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. 2008. Building a Sentiment Summarizer for Local Service Reviews. WWW Workshop



77

Baseline methods assume classes have equal frequencies!

- If not balanced (common in the real world)
  - precision not ideal for evaluation
  - need to use F-scores
- Severe imbalancing also can degrade classifier performance
- Common solutions:
  1. Resampling in training
     - Random undersampling
  2. Cost-sensitive learning
     - Penalize SVM more for misclassification of the rare thing
  3. Compare with dummy classifier (most common class)

78

Scherer Typology of Affective States

- **Emotion**: brief organically synchronized … evaluation of a major event
  - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes: enduring, affectively colored beliefs, dispositions towards objects or persons**
  - *liking, loving, hating, valuing, desiring*
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - *nervous, anxious, reckless, morose, hostile, jealous*

Computational work on other affective states

- **Emotion**:
  - Detecting annoyed callers to dialogue system
  - Detecting confused/frustrated versus confident students
- **Mood**:
  - Finding traumatized or depressed writers
- **Interpersonal stances**:
  - Detection of flirtation or friendliness in conversations
- **Personality traits**:
  - Detection of extroverts

Applications

COMFREAK@PIXABAY

Natural Language Generation

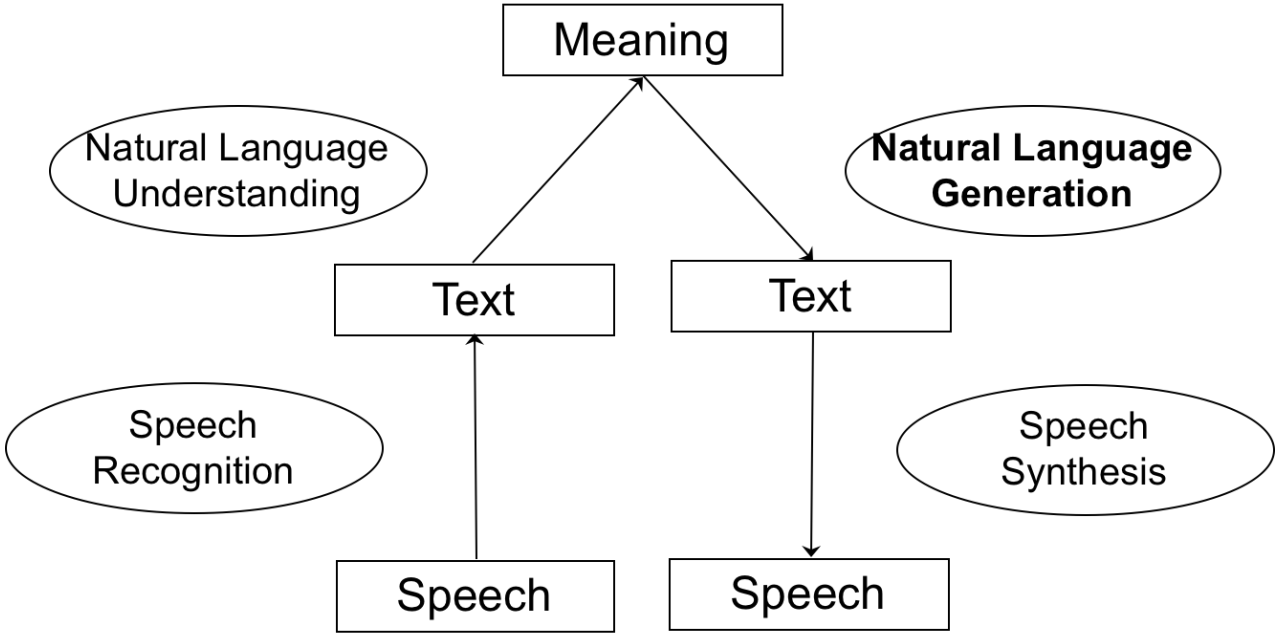
NLG systems are computer systems which produces understandable and appropriate texts in English or other human languages

- Input is data (raw, analysed)
- Output is documents, reports, explanations, help messages, and other kinds of texts

Requires

- Knowledge of language
- Knowledge of the domain

Natural language generation



Natural language generation -microplanning

Lexical/syntactic choice: Which words and linguistic structures to use?

Aggregation: How should information be distributed across sentences and paras

Reference: How should the text refer to objects and entities?

Question Answering (QA)

Question Answering: IBM's Watson

Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES
OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

86

Apple's Siri


AT&T
22:26
65 %
“ Do I need an umbrella tomorrow in San Francisco ”
Yes, San Francisco should get rain tomorrow:
Weekly Forecast
TUES WED THU FRI SAT SUN
14° 16° 17° 17° 18° 18°
9° 7° 6° 6° 7° 8°

Knowledge-based approaches (Siri)

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases

88

Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

89

This week

Semantics

- Word senses
- Path based similarity
- Information content similarity
- Distributional similarity

Applications

- Sentiment analysis
- Question answering
- NLG - Lexical choice

TU Delft

Next lecture

Evaluation of Natural Language Processing Systems

- Intrinsic (task-based) versus extrinsic (human ratings) evaluation
- Metric evaluations: BLEU, Rouge, METEOR
- Setting up statistical tests

Project proposal (due tomorrow)

Problem description: which problem will you tackle and what is interesting about the problem? If you reproduce a paper make sure to reference the original paper.
Resources: which data and tools will you be using
Methodology: if you choose to reproduce a paper, tell us which part of the paper (if not all) you will reproduce; if you choose your own research idea, tell us what type of algorithm/approach you are proposing and what your baseline(s) will be
Background readings: list at least 5 related papers that you will read to add context to your research. Reputable venues. NO ARXIV please!
Evaluation: how will you evaluate your algorithm/approach? Which evaluation metrics will you use? Need to be suitable for this task.

Submission: PDF, 1 per group. On Brightspace, should contain the group name and **group members** (name, student IDs).
Feedback: the course team will provide feedback on the project proposal within a few days.

Questions?