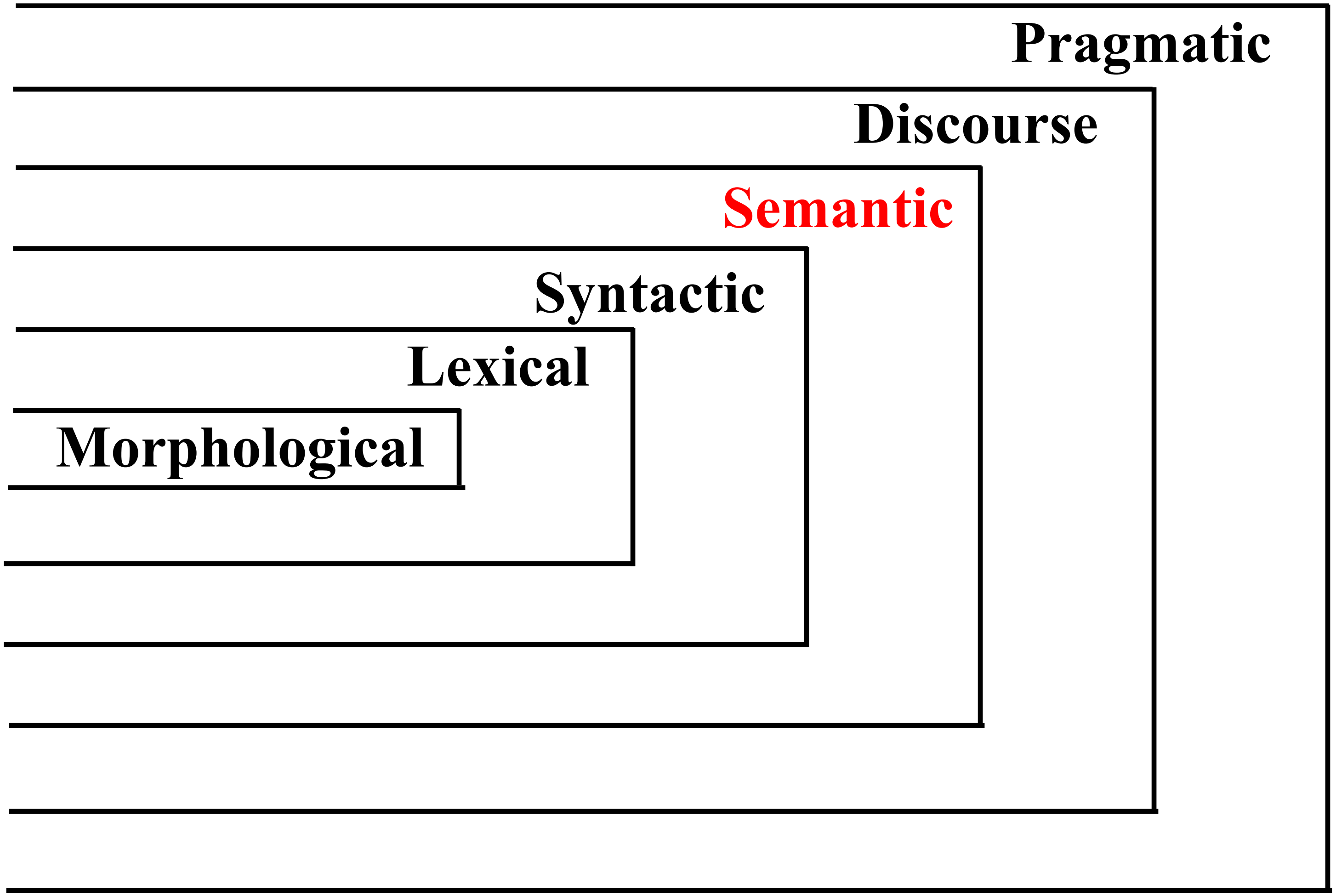
Semantic Processing, Semantic Representations, and Lexical Semantics:

Word Senses, WordNet, Ontologies
Word Sense Disambiguation
Semantic Lexical Resources
Topic Models (LDA)

With material from Miao Chen, Liz Liddy, Jurafsky and
Martin, and Rada Mihalcea

Synchronic Model of Language



A semantic theory:

- A theory of human ability to interpret the sentences of their language.
- Should predict whether a sentence is:
    - meaningful
    - ambiguous
    - anomalous

Interpretive vs. Generative Semantic Theories

- Semantic theories for "transformational generative syntax"
  - Syntax explaining how humans make well-formed sentences
- Interpretive semantic theories (Chomsky and Jackendoff) says that each syntactic structure can be assigned a meaning from a separate semantic theory
- Generative semantic theories (Katz and Fodor, "The structure of semantics", 1964) advocates a decompositional semantics that can build up the semantics of sentences using *semantic markers* as the interpretation of syntactic words and *selectional restrictions* on applying semantic relations.
  - The agent of the verb "kick" must be something active

Theories give rise to goals for semantic processing:

- Detect non-syntactic ambiguities. If a sentence is two ways ambiguous, characterize the meaning of each reading.

  The bill is large.

  - Eliminate ambiguities by using semantic relations within the sentence.

    The bill is large but I have enough money to cover it.
  - Related Topic: Word Sense Disambiguation
- Decide if one sentence is a paraphrase of another (two way).

  Your marks on the tests were excellent.

  You scored very high on the exams.
- Entailment: decide if the truth of one sentence implies the truth of another (one way).

  John lives in Toronto.
  implies *John's residence is in Canada.*

Relation between Syntax and Semantics in NLP

- Syntactic analysis:
  - determines the syntactic category of the words
  - assigns structural analysis to a sentence
  - what groups with what
- Semantic analysis:
  - Creation of a representation of the meaning of a sentence
- Clearly syntactic structure affects meaning (e.g. word order, phrase attachment).
  - *"The man with the telescope watched Mary."*
  - *"Mary watched the man with the telescope."*
- But meaning can determine syntactic structure

Syntax and Semantic Processing

- Process syntax first and then semantics
  - Intuitively appealing modularization
  - But some syntactic decisions not possible without semantics
- Process semantics first and syntax only as necessary
  - *"Without a full syntactic analysis, a system can miss possible meanings and accept impossible ones."* (Mitch Marcus)
- Process syntax and semantics as joint operations
  - Mainly guided by syntactic analysis
  - But partial semantics available when syntax needs guidance
- Note that Statistical Parsing at least uses some semantics in the form of word attachments or preferences in order to carry out the syntactic processing

Building blocks of semantic systems

- Semantics that words represent
  - Entities – individuals such as a particular person, location or product
    - John F. Kennedy, Washington, D.C., Cocoa Puffs
  - Concepts – the general category of individuals such as
    - person, city, breakfast cereal
  - Relations between entities and concepts
- Semantics indicated by verbs, prepositional phrases and other structures
  - Relations between concepts
    - Hierarchy of specific to more general concepts
    - Wide variety of other relations
  - Predicates representing verb structures
    - Semantic roles, case grammar

Semantic Representation

- A representation shows how to put together entities, concepts, relations and predicates to describe a situation or “semantic world”
  - Enables reasoning about that semantic world
- Some possible knowledge representation approaches:
  - First Order Logic
  - Semantic Nets
  - Conceptual Dependency
  - Frames
  - Rule-Based
  - Conceptual Graphs
  - Case Grammer

Why do we need semantic representations?

- To link the surface, linguistic elements to the non-linguistic knowledge of the world
  - Many words, few concepts
- Structures composed from a set of symbols
  - All languages have a predicate-argument structure
  - Correspond to relationships that hold among concepts underlying constituent words and phrases of a sentence, and then across sentences
- To represent the variety at the lexical level at a unified conceptual level
  - Unambiguous representations; canonical forms
- Can be used to reason, both to verify what is true in the world and to infer knowledge from the semantic representation

First Order Logic

- Also known as Predicate Calculus

- A symbolic language whose symbols have precisely stated meanings and uses
  - The symbols can be used as meanings in the real world
  - Typically express properties of entities in the world

- First Order Logic (FOL) often used in AI systems found in such applications as robotics and computational control systems
  - Allows a natural language interface to such systems

- Example – *if Socrates is a man, then Socrates is a mortal*

  Man (Socrates) -> Mortal (Socrates)

FOL language

- FOL uses terms to represent objects in the real world
  - Constants are specific objects in the world - entities
    - Socrates, Pastabilities
  - Functions represent concepts about objects
    - LocationOf (Pastabilities)
  - Variables are used to stand for any object
    - X
- FOL uses predicates to state relations between objects
  - If Serves is a predicate taking a restaurant and a type of food as arguments, we can state that a particular restaurant serves a type of food
    - Serves (Pastabilities, VegetarianFood)

FOL language, cont.

- FOL uses connectives *and* and *or* to combine statements
  - Serves (Pastabilites, VegetarianFood) ^ IsExpensive(Pastabilities)
- FOL uses the implication connection to mean if the first statement is true, then the second one is also true
  - Serves(Pastabilities, VegetarianFood) => Restaurant(Pastabilities)
  - Is this true?
- FOL uses the existential quantifier to assert that an object with particular properties exists

  ∃ x Restaurant(x) ^ Serves(x, VegetarianFood)
- FOL uses the universal quantifier to assert that particular properties are true for all objects
  - “ (forall) x Restaurant(x) => Serves(x, VegetarianFood)

  (this is definitely false because not all restaurants serve vegetarian food)

Example

A person born in the United Kingdom after commencement shall be a British citizen if at the time of birth his father or mother is:

a. a British citizen; or

b. settled in the United Kingdom

Which can be represented as:

```
(forall x)(there exists y, z)
    ((x was born in the U.K.)
    ^ (x was born on date y
    ^ (y is after or on commencement)
    ^ (z is a parent of x)
    ^ (z is a British citizen on date y))
    => (x is a British citizen)
```

Reasoning with FOL

- FOL allows inference to make conclusions of new information
  - Inference rule is called “modus ponens”, informally is if-then reasoning

    if we know that A is true and we know that A => B is true,
    we can conclude that B is true

Events in First Order Logic

- So far the predicates have captured state, properties that remain unchanged over some period of time
- Events denote changes in some state and can have a host of participants, props, times and locations.
- One way to give events in FOL is to state the existence of an event that has all the participants, etc.

I ate a turkey sandwich for lunch at my desk on Tuesday.

∃ e Eating(e) ^ Eater(e, Speaker) ^ Eaten(e, TurkeySandwich)
^ Meal(e, Lunch) ^ LocationOf(e, Desk) ^ Time(e, Tuesday)

Difficulties with First Order Logic

- Problem for NLP:
  - ‘semantics’ of logic does not necessarily equate to ‘meaning’ in the real world
  - Not everything is as clear cut as required by a formal logic
  - May not be enough “real world” predicates in the FOL system to capture semantics of text
    - This is a problem for all the semantic representations
    - Semantic systems better developed for objects and actions
    - Not as well developed to represent ideas and beliefs
  - See Cyc Corp efforts to embody all world knowledge in (essentially) First Order Logic
    - http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc

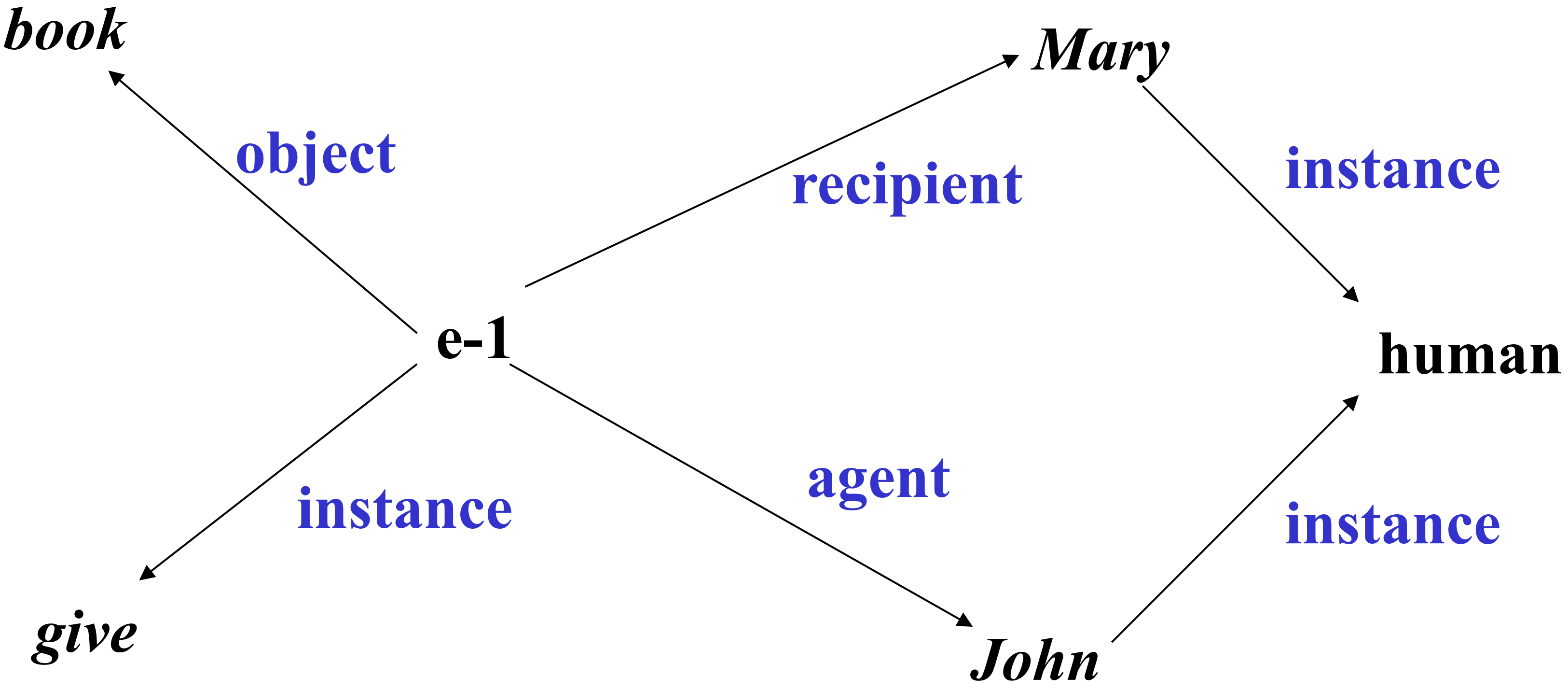
In-class exercise.

Semantic Networks

A network or graph of nodes joined by links where:

- nodes represent concepts (e.g. BOOK, GREEN)
- links (labelled, directed arcs) represent relations (e.g. ISA)

John gives a book to Mary.



Frames

- A type of structured representation or *schema*
- Introduced by Marvin Minsky in 1975
  - "A Framework for Representing Knowledge"
  - Most widely referenced paper on knowledge representation
  - Explicitly attempts to represent human processing
- Based on common sense knowledge
- A way of grouping information about an entity or an event or a state in terms of a record of 'slots' and 'fillers'
  - One slot filled by the name of the object that the node stands for
  - Other slots filled with value of various common attributes associated with such an object
    - These can be properties or relations

Example of Frames

- Wikipedia Info Box is an example of a frame structure
  - Slot names are attributes or relations
  - An attribute value is information such as a date or height
  - A relation value is another entity, which may have its own frame
- More formal frame systems require uniformity of slot names and value syntax

Name	Barack Obama
Birthdate	August 4, 1961
Birthplace	Honolu, Hawaii
Height	6’ 1” (1.85 m)
Parents	Barack Obama Sr., Ann Dunham
Children	Natasha Obama, Malia Ann Obama

Applications of Semantic Representations

- Semantic representations are used to represent entities with their properties and relations in information extraction and question answering systems
- Semantic representations are used in reasoning in AI systems such as robot manipulations
  - Could also be used in dialog systems
  - Works best in a small environment where the amount of world knowledge needed is small

Getting Semantic Representation from Text

- Use a syntactic parse tree to identify predicates and possible relations structures
- Algorithms map syntactic structure to relations, given the words in the text
  - Semantic role labeling is one important algorithm (next section)
  - Some systems employ a First Order Logic mapper
  - Watson (IBM question answering system) mapped dependency parses to frames

Lexical Semantics

- Lexemes – individual entries in a lexicon
  - Senses apply to lexemes, which are some form of the root word rather than orthographic form
- In recent years, most dictionaries made available in Machine Readable format (MRD)
  - Many online dictionaries
- Thesauruses – add synonymy information
  - Roget Thesaurus
- Semantic networks – add more semantic relations
  - WordNet
  - EuroWordNet
- Ontologies – add semantic relations and rules about entities and concepts

WordNet

- WordNet is a database of facts about words
  - Meanings and the relations among them
- Words are organized into clusters of synonyms
  - Synsets
- http://wordnet.princeton.edu/
- Organized into nouns, verbs, adjectives, and adverbs
  - Currently 170,000 synsets
  - Available for download, arranged in separate files (DBs)

MRD – Knowledge Resources

- For each word in the language vocabulary, an MRD provides:
  - A list of meanings
  - Definitions (for all word meanings)
  - Typical usage examples (for most word meanings)

WordNet definitions(called glosses)/examples for synsets of the noun *plant*

1. buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles“
2. a living organism lacking the power of locomotion
3. something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"
4. an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

MRD – Knowledge Resources

- A thesaurus adds:
  - An explicit synonymy relation between word meanings

WordNet synsets for the noun “plant”

1. plant, works, industrial plant
2. plant, flora, plant life

- A semantic network adds relations:
  - Hypernymy/hyponymy (IS-A), meronymy/holonymy (PART-OF), antonymy, entailnment, etc.

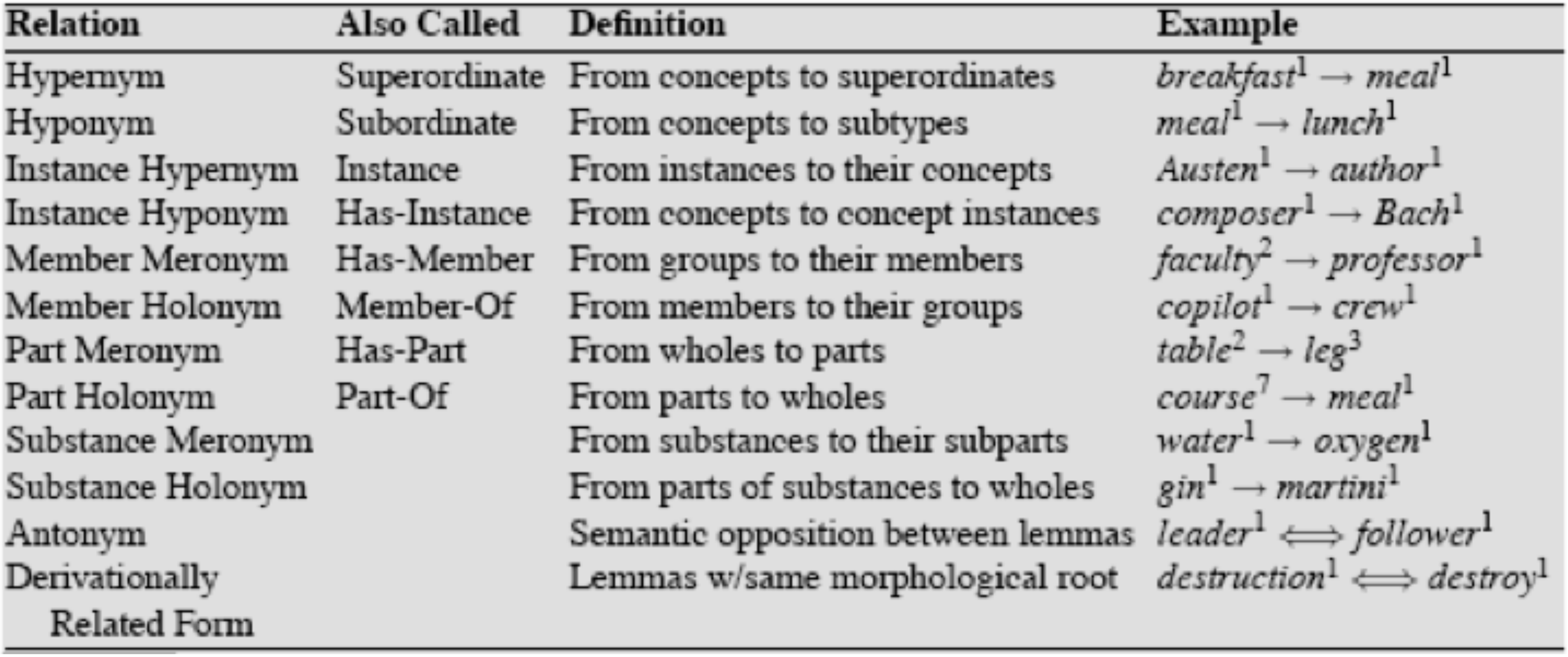
WordNet related concepts for the meaning “plant life”

{plant, flora, plant life}

hypernym:	{organism, being}
hypomym:	{house plant}, {fungus}, …
meronym:	{plant tissue}, {plant part}
holonym:	{Plantae, kingdom Plantae, plant kingdom}

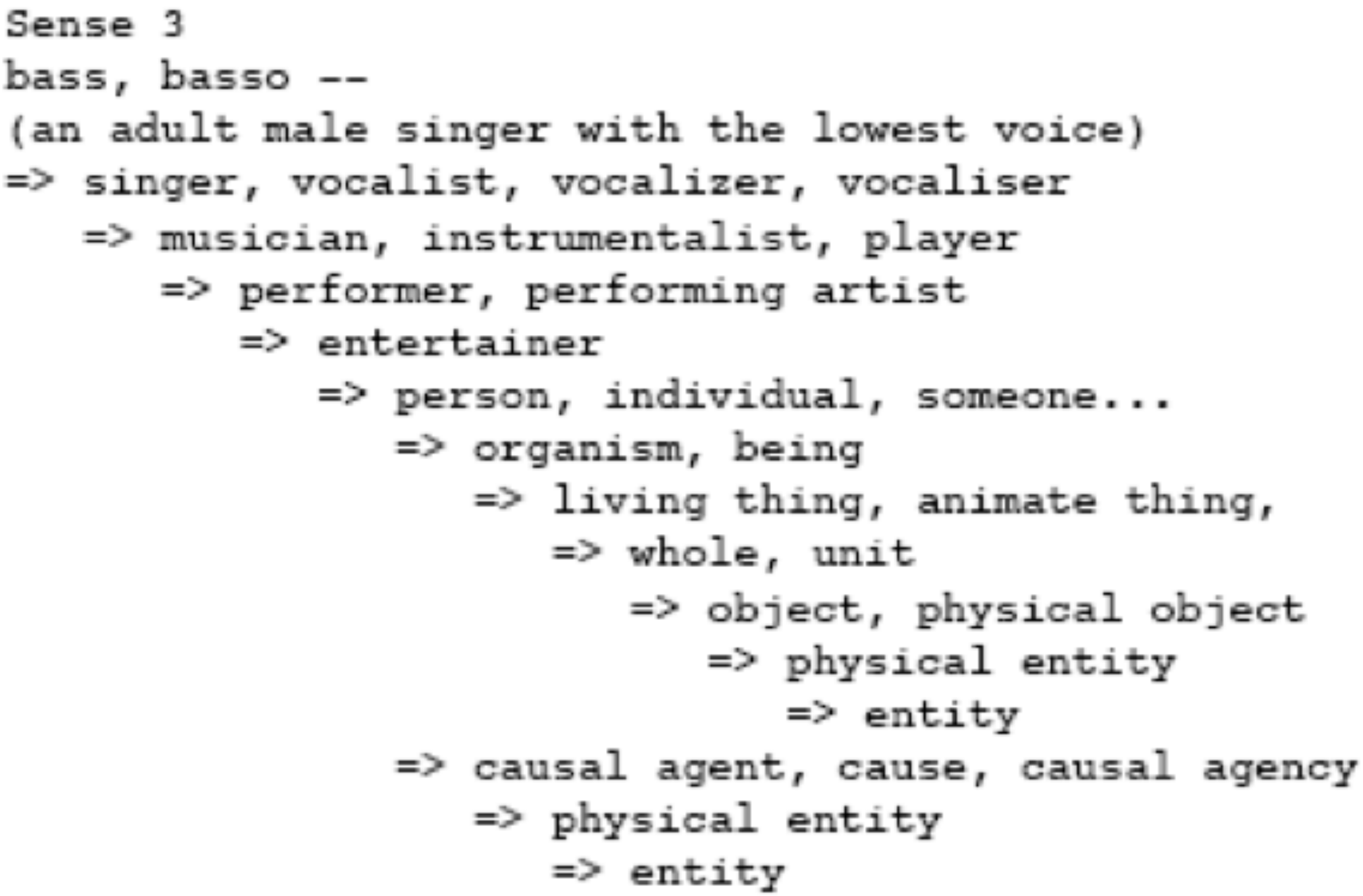
WordNet Relations

- A more detailed list from Jurafsky and Martin



Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} → *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} → *lunch*^{1}
Instance Hypernym	Instance	From instances to their concepts	*Austen*^{1} → *author*^{1}
Instance Hyponym	Has-Instance	From concepts to concept instances	*composer*^{1} → *Bach*^{1}
Member Meronym	Has-Member	From groups to their members	*faculty*^{2} → *professor*^{1}
Member Holonym	Member-Of	From members to their groups	*copilot*^{1} → *crew*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} → *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} → *meal*^{1}
Substance Meronym		From substances to their subparts	*water*^{1} → *oxygen*^{1}
Substance Holonym		From parts of substances to wholes	*gin*^{1} → *martini*^{1}
Antonym		Semantic opposition between lemmas	*leader*^{1} ⟺ *follower*^{1}
Derivationally Related Form		Lemmas w/same morphological root	*destruction*^{1} ⟺ *destroy*^{1}

WordNet Hierarchies



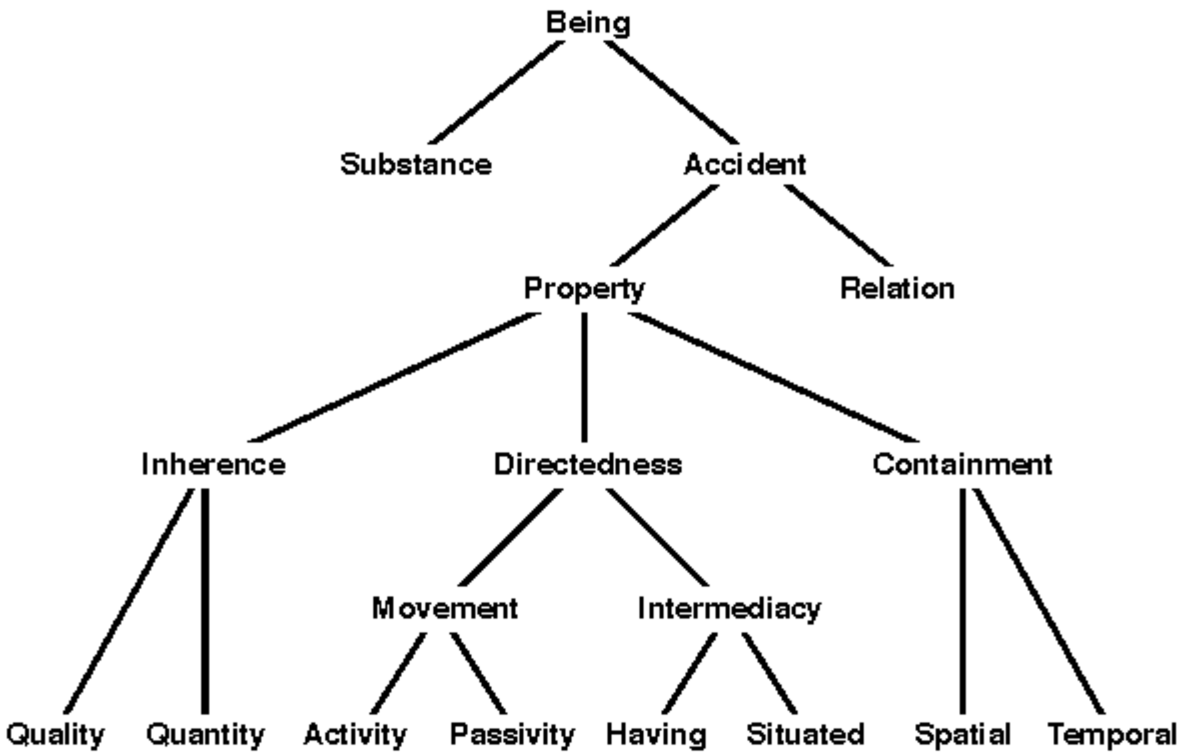
```
Sense 3
bass, basso --
(an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser
   => musician, instrumentalist, player
      => performer, performing artist
         => entertainer
            => person, individual, someone...
               => organism, being
                  => living thing, animate thing,
                     => whole, unit
                        => object, physical object
                           => physical entity
                              => entity
               => causal agent, cause, causal agency
                  => physical entity
                     => entity
```

What is Ontology?

- In philosophy, ontology studies existence/being of the world.
- A fundamental question: What is there?

  Answer: Everything
- It can be construed that ontology is about categorizing existence/being of the world.
- An example: Aristotle’ s ontology

Aristotle's Ontology



- In his work "categories", Aristotle listed ten categories that all things of the world should belong to.

http://www.jfsowa.com/talks/ontology.htm

Ontology in Information Science

- Ontology is an approach of knowledge representation.
- Ontology is a specification of a conceptualization (Gruber, 1993).
- Major components:

Concepts, i.e. *human, animal, food, table, movie, etc*.

Instances, i.e. Miao Chen is an instance of concept "person".

Properties, i.e. a human has properties of *gender, height, weight, father, mother, etc.*

Relations, i.e. Syracuse University is *located in* Syracuse.

Rules. If someone is married, then he/she should have a spouse.

- Given some text, how would you represent its knowledge in ontology?

Ontology to represent semantics

- Barack Hussein Obama II (born August 4, 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama was the junior United States Senator from Illinois from January 2005 until November 2008, when he resigned following his election to the presidency. (http://en.wikipedia.org/wiki/Obama)

Relations: is-a, resign, is-president-of

Barack Obama has properties of

Gender: male

Race: African American

Administrative role: President of the United States

Ontology Examples: UMLS and WordNet

- The Unified Medical Language System (UMLS) aggregates various controlled vocabularies and mapped them to a comprehensive biomedical ontology. It has three knowledge sources:

  Metathesaurus. Mapping concepts and terms in different thesaurus and organizing them in the UMLS structure

  Semantic network. Connecting semantic types of concepts in metathesaurus by semantic relations.

  Specialist Lexicon. Containing lexical information of biomedical terms.

  http://umlsks.nlm.nih.gov/uPortal/render.userLayoutRootNode.uP

  (registered users only)

- The WordNet, a lexical database, can be viewed a light-weighted ontology. The resource includes nouns, verbs, adjectives, and adverbs, with synonymous terms grouped together (in synsets).
- Look at WordNet online viewer: http://wordnet.princeton.edu/

Word Sense Disambiguation

- Definition
  - Correct selection of the appropriate sense / meaning of a polysemous word in context
- In English, the most frequently occurring nouns have 7 senses and the most frequently occurring verbs have 11 senses
- How can we define different word senses?
  - Give a list of synonyms
  - Give a definition, which will necessarily use words that will have different senses, and these will (perhaps circularly) use words for definitions
- Coarse-grained senses distinguish core aspects of meaning
- Fine-grained senses also distinguish periphal aspects of meaning

Difficulties with synonyms

- True synonyms non-existent, or very rare
- Near-synonyms (Edmonds and Hirst)
  - Examples:
    - Error, blunder, mistake
    - Order, command, bid, enjoin, direct
  - Dimensions of synonym differentiation
    - Stylistic variation
      - Pissed, drunk, inebriated
    - Expressive variation
      - Attitude: skinny, thin, slim
      - Emotion: father, dad, daddy
    - . . .

Approaches

- Sense Inventory usually comes from a dictionary or thesaurus.
- Progression of approaches
  - 1970s - 1980s
    - Rule based systems
    - Rely on hand crafted knowledge sources
  - 1990s
    - Corpus based approaches
    - Dependence on sense tagged text
  - 2000s
    - Hybrid Systems
    - Minimizing or eliminating use of sense tagged text
    - Taking advantage of the Web
- Reasonable to consider how humans do it

Human Sense Disambiguation

- Sources of influence known from psycholinguistics research:
  - local context
    - the sentence containing the ambiguous word restricts the interpretation of the ambiguous word
  - domain knowledge
    - the fact that a text is concerned with a particular domain activates only the sense appropriate to that domain
  - frequency data
    - the frequency of each sense in general usage affects its accessibility to the mind

Lesk Algorithm

- Original Lesk definition: measure overlap between sense definitions for all words in context. (Michael Lesk 1986)
  - Identify simultaneously the correct senses for all words in context
- Simplified Lesk (Kilgarriff & Rosensweig 2000): measure overlap between sense definitions of a word and current context
  - Identify the correct sense for one word at a time
  - Current context is the set of words in the surrounding sentence/paragraph/document.

Lesk Algorithm: A Simplified Version

- Algorithm for simplified Lesk:

  1.Retrieve from MRD all sense definitions of the word to be disambiguated

  2.Determine the overlap between each sense definition and the current context

  3.Choose the sense that leads to highest overlap

Example: disambiguate PINE in

"Pine cones hanging in a tree"

- PINE

  1. kinds of evergreen tree with needle-shaped leaves
  2. waste away through sorrow or illness

Pine#1 ∩ Sentence = 1
Pine#2 ∩ Sentence = 0

Evaluations of Lesk Algorithm

- Initial evaluation by M. Lesk
  - 50-70% on short samples of text manually annotated set, with respect to Oxford Advanced Learner's Dictionary
  - Set of senses are "coarse-grained"
- Senseval evaluation conferences have shared tasks involving data for word sense disambiguation
  - Uses WordNet senses (more fine-grained and thus more difficult)
  - Evaluation on Senseval-2 all-words data, with back-off to random sense (Mihalcea & Tarau 2004)
    - Original Lesk: 35%
    - Simplified Lesk: 47%
  - Evaluation on Senseval-2 all-words data, with back-off to most frequent sense (Vasilescu, Langlais, Lapalme 2004)
    - Original Lesk: 42%
    - Simplified Lesk: 58%

WSD algorithm development in Senseval

- Lexical sample task
  - Small pre-selected set of target words
  - Inventory of senses for each word from some lexicon
  - Various labeled corpora developed for each word
  - Suitable for specific domain applications with small number of words
- All-word task
  - Given an entire text, disambiguate every content word in the text
  - Use general-purpose lexicon with senses
  - Can use a labeled corpus
    - SemCor is a subset of the Brown corpus with 234,000 words labeled with WordNet senses
    - Additional corpora developed through Senseval

Sense Tagged Corpus

- Examples of sense tagged text:

Bonnie and Clyde are two really famous criminals, I think they were **bank/1** robbers
My **bank/1** charges too much for an overdraft.
I went to the **bank/1** to deposit my check and get a new ATM card.
The University of Minnesota has an East and a West **Bank/2** campus right on the Mississippi River.
My grandfather planted his pole in the **bank/2** and got a great big catfish!
The **bank/2** is pretty muddy, I can’t walk there.

Classification approach to WSD

- Often referred to as Supervised Learning approach
- Train a classification algorithm that can label each (open-class) word with the correct sense, given the context of the word
- Training set is the hand-labeled corpus of senses
- The context is represented as a set of “features” of the word and includes information about the surrounding words
- Result of training is a model that is used by the classification algorithm to label words in the test set, and ultimately, in new text examples

WSD classification features

- Collocational features
  - Information about words in specific positions (i.e. previous word)
  - Typical features include the word itself, its stem and its POS tag
  - Example feature set:

    2 words to the left and right of the target word and their POS tags

    An electric guitar and ***bass*** *player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps.*

    [guitar, NN, and, CC, player, NN, stand, VB]
- Syntactic features
  - Predicate-argument relations
    - Verb-object, subject-verb,
  - Heads of Noun and Verb Phrases

WSD classification features

- Bag-of-words features
  - Unordered set of words with position ignored from context
  - Context is typically small fixed-size window.
  - Context words may be limited to a small number of frequently-used context words.
  - Example: for each word, collect the 12 most frequent words from a collection of sentences drawn from the corpus as the limited set.

    For bass, the 12 most frequent context words from the WSJ are:

    [fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band]

    The features of bass in the previous sentence (represented as 1 or 0 indicating the presence or not of the word in a window of size 10):

    [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0]

Results for supervised learning systems

- Accuracy of different systems applied to the same data tends to converge on a particular value, no one system shockingly better than another.
  - Senseval-1, a number of systems in range of 74-78% accuracy for English Lexical Sample task.
  - Senseval-2, a number of systems in range of 61-64% accuracy for English Lexical Sample task.
  - Senseval-3, a number of systems in range of 70-73% accuracy for English Lexical Sample task…
- What to do next?
  - Difficulty of creating enough annotated data to obtain an accurately trained classifier

Semi-supervised Classification Approaches

- Requires:
  - A small amount of annotated text
  - A large amount of plain unannotated text
  - A way to determine if a labeled example is most likely correct
- Approach:
  - Train a classifier on the annotated text
  - Run it on the unannotated text to label word senses
  - For every labeled example that is most likely correct, add it to the annotated text
  - Repeat until no more most likely correct examples are achieved
- Unannotated Corpus
  - Can be a pre-defined collection
  - Can be generated from the web by formulating queries with contextual clues

WSD algorithms in applications

- Information retrieval:
  - *Example query: I would like information about developments in low-risk instruments, especially those being offered by companies specializing in bonds.*
  - Try to improve retrieval results by using WSD to find the correct sense of each word and add synonyms to the expanded query
  - Results have not been very successful
- Machine Translation
  - WSD has been successful in improving the correct translations

Semantic classes for words: Wiebe et al

- Lexical resources have been developed to assign words to semantic classes in support of applications that need to detect opinion, sentiment, or other more subjective meanings
- Subjectivity Lexicon by Wiebe et al
  - Gives a list of 8,000+ words that have been judged to be weakly or strongly positive, negative or neutral in subjectivity
  - Examples:

```
type=weaksubj len=1 word1=abandoned pos1=adj stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=abandonment pos1=noun stemmed1=n priorpolarity=negative
type=weaksubj len=1 word1=abandon pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abase pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abasement pos1=anypos stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=abash pos1=verb stemmed1=y priorpolarity=negative
type=weaksubj len=1 word1=abate pos1=verb stemmed1=y priorpolarity=negative
type=strongsubj len=1 word1=absolve pos1=verb stemmed1=y priorpolarity=positive
type=strongsubj len=1 word1=absolute pos1=adj stemmed1=n priorpolarity=neutral
```

Semantic classes for words: LIWC

- Linguistic Inquiry and Word Count
  - Text analysis software based on dictionaries of word dimensions
  - Dimensions can be syntactic
    - Pronouns, past-tense verbs
  - Dimensions can be semantic
    - Social words, affect, cognitive mechanisms
  - Other categories
    - See http://www.liwc.net/comparedicts.php
- James Pennebaker, Univ. of Texas at Austin
  - http://www.liwc.net/
- Often used for positive and negative emotion words in opinion mining

Semantic classes for words: ANEW

- Affective Norms for English Words
  - Provides a set of emotional ratings for a large number of words in the English language
- Participants gave graded reactions from 1-9 on three dimensions
  - Good/bad, psychological valence
  - Active/passive, arousal valence
  - Strong/weak, dominance valence
- From the NIMH Center for the Study of Emotion and Attention at the University of Florida
  - http://csea.phhp.ufl.edu/Media.html
  - See also the paper by Dodds and Danforth on Happiness of Large-Scale Written Expressions

Topic Models

- LDA (Latent Dirichlet Allocation) follows other research in
  - going beyond unigram models, where word presence or frequency is considered independently of other words
  - to finding intra-document statistical structure that can reveal a model of word co-occurrence within the document
    - Blei, Ng, Jordan “Latent Dirichlet Allocation”, Journal of Machine Learning Research, 2003.
- Related idea to LSI (Latent Semantic Indexing)
  - Used mathematical techniques to reduce the dimensionality of the word-document occurrence matrix used in Information Retrieval
    - Revealed semantic similarities between words due to their co-occurrence within documents (synomous)
    - Revealed possible semantic dissimilarities of the same word (polysonomous)

Topic Models

- Topic model
  - a collection of words together with the probabilities that they are used in that particular topic
- Generative model for documents
  - A document is generated as a mixture of topics, with words used according to the probability distribution of the topic and to the percentage that the topic contributes to the mixture
- Problem of statistical inference
  - Given the words of a document, find the topic model that is most likely to have generated those words
    - Find the probability distribution of words in each topic
    - Find the distribution of topics over the document

Documents as mixture of topics

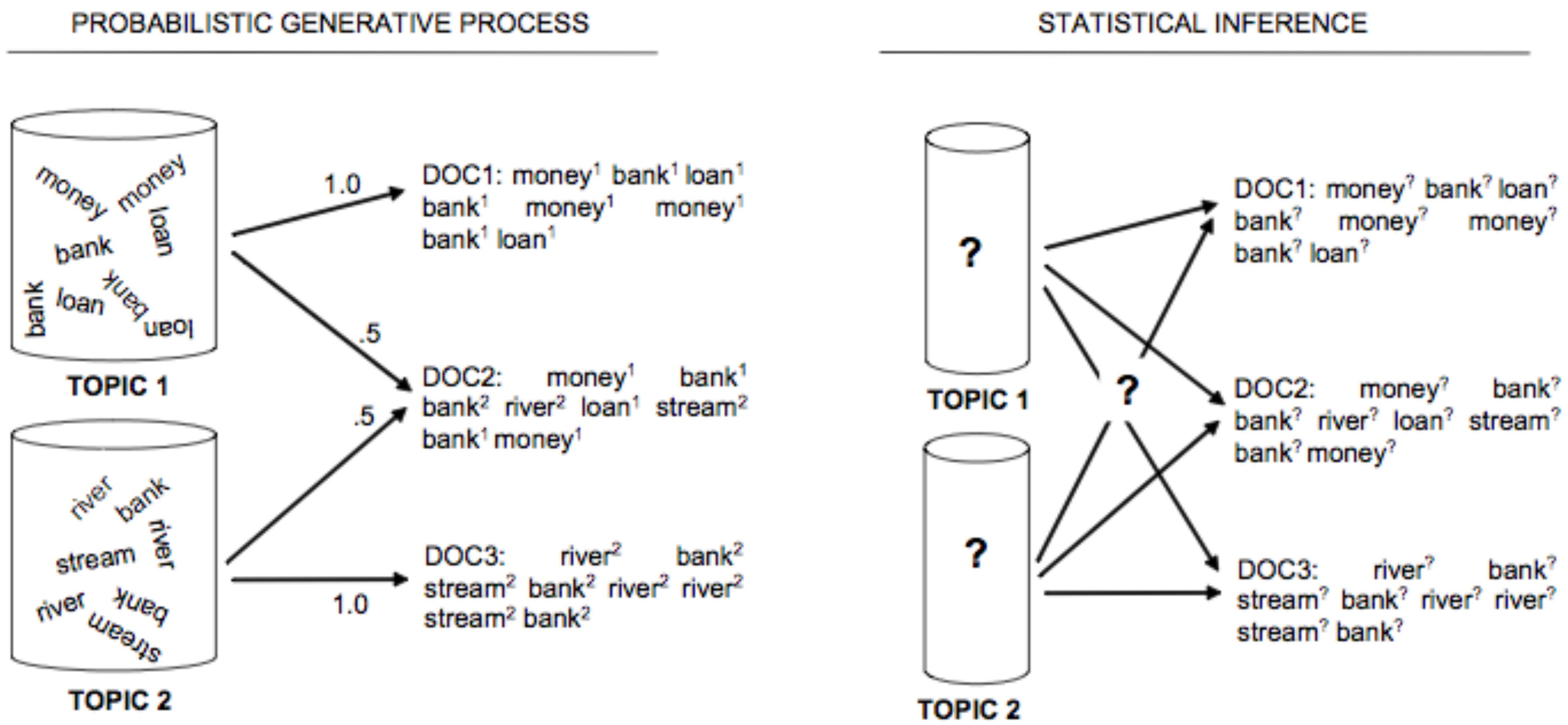
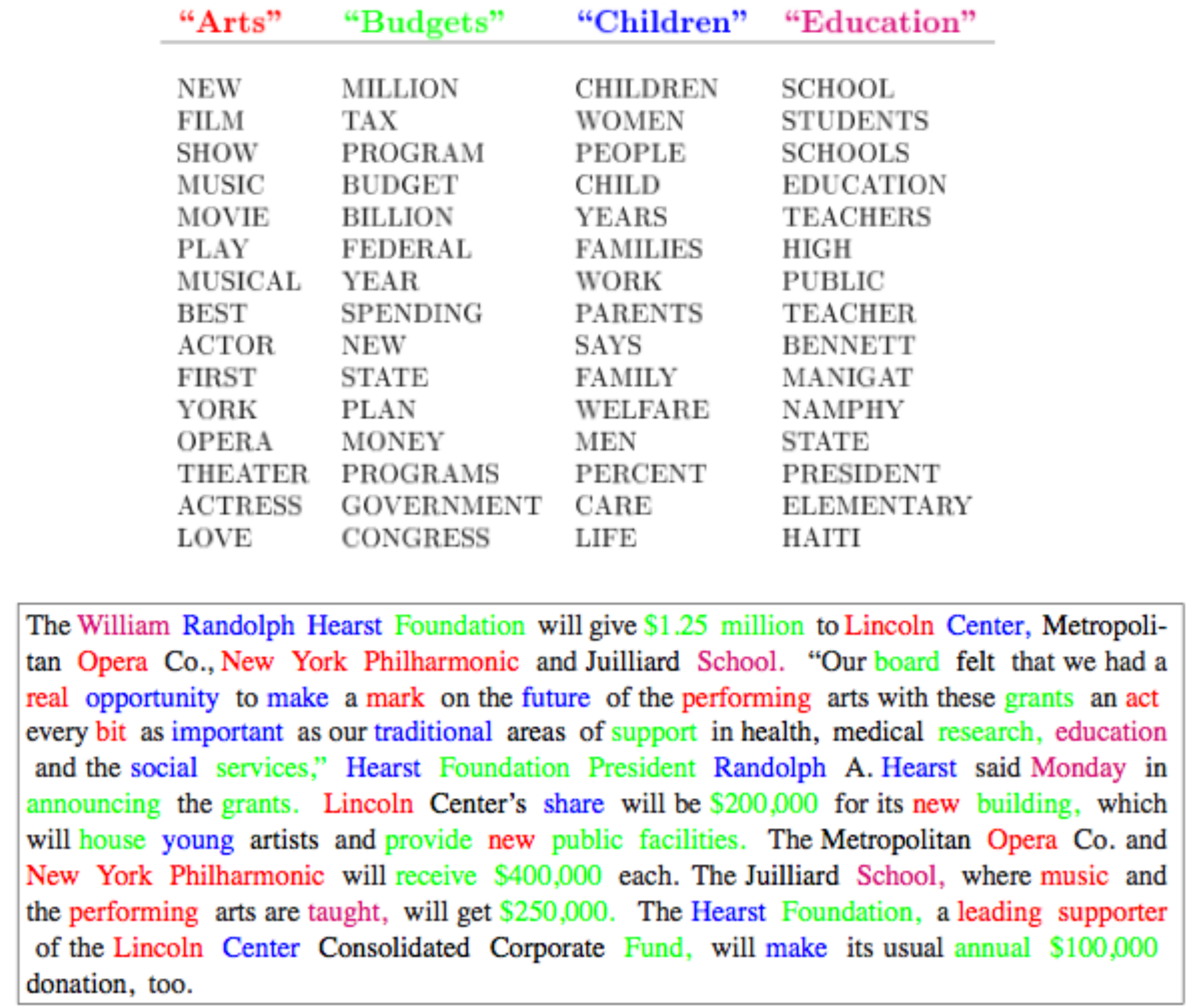

PROBABILISTIC GENERATIVE PROCESS
STATISTICAL INFERENCE
money money loan bank bank loan bank loan
TOPIC 1
river bank river stream river bank stream
TOPIC 2
1.0
.5
.5
1.0
DOC1: money^1 bank^1 loan^1 bank^1 money^1 money^1 bank^1 loan^1
DOC2: money^1 bank^1 bank^2 river^2 loan^1 stream^2 bank^1 money^1
DOC3: river^2 bank^2 stream^2 bank^2 river^2 river^2 stream^2 bank^2
?
TOPIC 1
?
?
TOPIC 2
DOC1: money$^?$ bank$^?$ loan$^?$ bank$^?$ money$^?$ money$^?$ bank$^?$ loan$^?$
DOC2: money$^?$ bank$^?$ bank$^?$ river$^?$ loan$^?$ stream$^?$ bank$^?$ money$^?$
DOC3: river$^?$ bank$^?$ stream$^?$ bank$^?$ river$^?$ river$^?$ stream$^?$ bank$^?$

Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

From Steyvers and Griffiths “Probabilistic Topic Models”

LDA applications

- LDA software infers topic models from a document collection
- Typically, you must specify how many topics you want (and possibly experiment to get satisfactory ones)
  - Also may be necessary to tune parameters alpha and beta
- Topics are represented as lists of words, sometimes with the probabilities of the topic model
  - Note that in the example on the next slide, humans added the overall "topic" heading at the top
  - Software available form
    - Blei
    - Mallet
    - Stanford



"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

From Blei et al