Statistical Models of Semantics and Unsupervised Language Discovery

Andrew McCallum

Computer Science Department
University of Massachusetts Amherst

Including slides from Chris Manning and Dan Klein.

Attachment Ambiguity

- Where to attach a phrase in the parse tree?
- *"I saw the man with the telescope."*
  - What does "with a telescope" modify?
  - Is the problem AI complete? Yes, but…

  - Proposed simple structural factors
    - Right association [Kimball 1973]
      'low' or 'near' attachment = 'early closure' of NP
    - Minimal attachment [Frazier 1978]
      (depends on grammar) = 'high' or 'distant' attachment
      = 'late closure' (of NP)

Attachment Ambiguity

- “The children ate the cake with a spoon.”
- “The children ate the cake with frosting.”

- “Joe included the package for Susan.”
- “Joe carried the package for Susan.”

- *Ford, Bresnan and Kaplan (1982):*
  “It is quite evident, then, that the closure effects in these sentences are induced in some way by the choice of the lexical items.”

Lexical acquisition, semantic similarity

- Previous models give same estimate to all unseen events.
- Unrealistic - could hope to refine that based on semantic classes of words
- Examples
  - “Susan ate the cake with a durian.”
  - “Susan had never eaten a fresh durian before.”
  - Although never seen “eating pineapple” should be more likely than “eating holograms” because pineapple is similar to apples, and we have seen “eating apples”.

An application: selectional preferences

- Most verbs prefer arguments of a particular type. Such regularities are called *selectional preferences* or *selectional restrictions*.
- "Bill drove a…"   Mustang, car, truck, jeep

- Selectional preference strength: how strongly does a verb constrain direct objects
- "see" versus "unknotted"

Measuring selectional preference strength

- Assume we are given a clustering of (direct object) nouns. Resnick (1993) uses WordNet.

- Selectional association between a verb and a class

$$S(v) = D(P(C|v)||P(C) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)}$$

Proportion that its summand contributes to preference strength.

$$A(v,c) = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{S(v)}$$

- For nouns in multiple classes, disambiguate as most likely sense: $A(v,n) = \max_{c \in \text{classes}(n)} A(v,c)$

Selection preference strength (made up data)

Noun class c	P(c)	P(c|eat)	P(c|see)	P(c|find)
people	0.25	0.01	0.25	0.33
furniture	0.25	0.01	0.25	0.33
food	0.25	0.97	0.25	0.33
action	0.25	0.01	0.25	0.01
SPS S(v)		**1.76**	**0.00**	**0.35**

A(eat, food) = 1.08

A(find, action) = -0.13

Selectional Preference Strength example
(Resnick, Brown corpus)

Verb v	Noun n	$A(v, n)$	Class	Noun n	$A(v, n)$	Class
answer	*request*	4.49	speech act	*tragedy*	3.88	communication
find	*label*	1.10	abstraction	*fever*	0.22	psych. feature
hear	*story*	1.89	communication	*issue*	1.89	communication
remember	*reply*	1.31	statement	*smoke*	0.20	article of commerce
repeat	*comment*	1.23	communication	*journal*	1.23	communication
read	*article*	6.80	writing	*fashion*	−0.20	activity
see	*friend*	5.79	entity	*method*	−0.01	method
write	*letter*	7.26	writing	*market*	0.00	commerce

But how might we measure word similarity for word classes?

- Vector spaces

A document-by-word matrix A.

	cosmonaut	astronaut	moon	car	truck
d_1	1	0	1	1	0
d_2	0	1	1	0	0
d_3	1	0	0	0	0
d_4	0	0	0	1	1
d_5	0	0	0	1	0
d_6	0	0	0	0	1

But how might we measure word similarity for word classes?

- Vector spaces

  word-by-word matrix B

	cosmonaut	astronaut	moon	car	truck
cosmonaut	2	0	1	1	0
astronaut	0	1	1	0	0
moon	1	1	2	1	0
car	1	0	1	3	1
truck	0	0	0	1	2

A modifier-by-head matrix C

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

Similarity measures for binary vectors

Similarity measure	Definition
matching coefficient	$\|X \cap Y\|$
Dice coefficient	$\frac{2\|X \cap Y\|}{\|X\| + \|Y\|}$
Jaccard coefficient	$\frac{\|X \cap Y\|}{\|X \cup Y\|}$
Overlap coefficient	$\frac{\|X \cap Y\|}{\min(\|X\|, \|Y\|)}$
cosine	$\frac{\|X \cap Y\|}{\sqrt{\|X\| \times \|Y\|}}$

Cosine measure

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

maps vectors onto unit circle by dividing through

by lengths:

$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Example of cosine measure on word-by-word matrix on NYT

Focus word	Nearest neighbors							
garlic	*sauce*	.732	*pepper*	.728	*salt*	.726	*cup*	.726
fallen	*fell*	.932	*decline*	.931	*rise*	.930	*drop*	.929
engineered	*genetically*	.758	*drugs*	.688	*research*	.687	*drug*	.685
Alfred	*named*	.814	*Robert*	.809	*William*	.808	*W*	.808
simple	*something*	.964	*things*	.963	*You*	.963	*always*	.962

Probabilistic measures

(Dis-)similarity measure	Definition
KL divergence	$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$
Skew	$D(q\|\alpha r + (1-\alpha)q)$
Jensen-Shannon (was IRad)	$\frac{1}{2}D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2})$
L_1 norm (Manhattan)	$\sum_i \lvert p_i - q_i \rvert$

Neighbors of word "company"
[Lee]

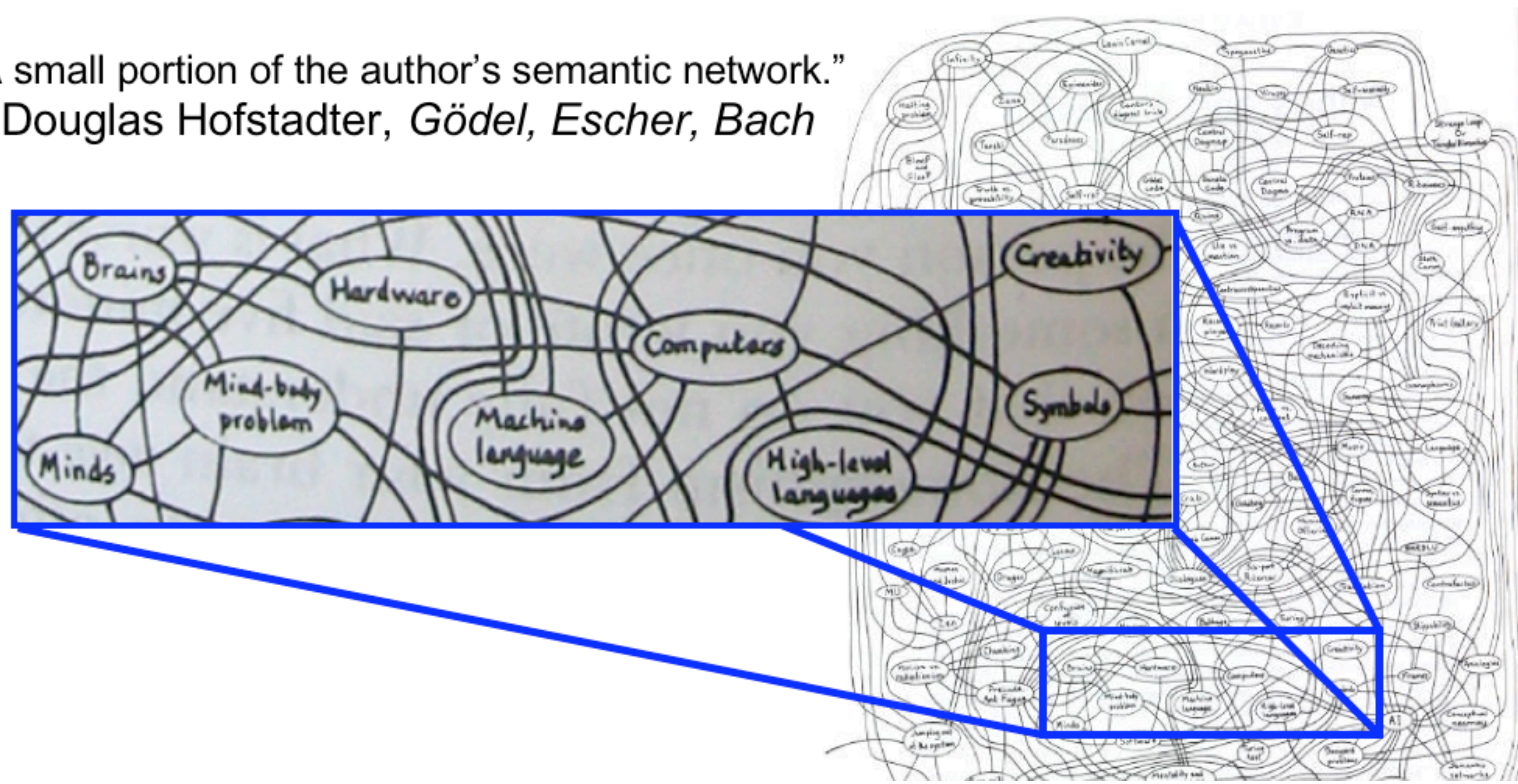
Skew ($\alpha = 0.99$)	J.-S.	Euclidean
airline	business	city
business	airline	airline
bank	firm	industry
agency	bank	program
firm	state	organization
department	agency	bank
manufacturer	group	system
network	govt.	today
industry	city	series
govt.	industry	portion

Learning syntactic patterns for automatic hypernym discovery

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng.

- It has long been a goal of AI to automatically acquire structured knowledge directly from text, e.g, in the form of a semantic network.

“A small portion of the author’s semantic network.”
– Douglas Hofstadter, *Gödel, Escher, Bach*



We aim to classify whether a noun pair (X, Y) participates in one of the following semantic relationships:

Hypernymy (ancestor)

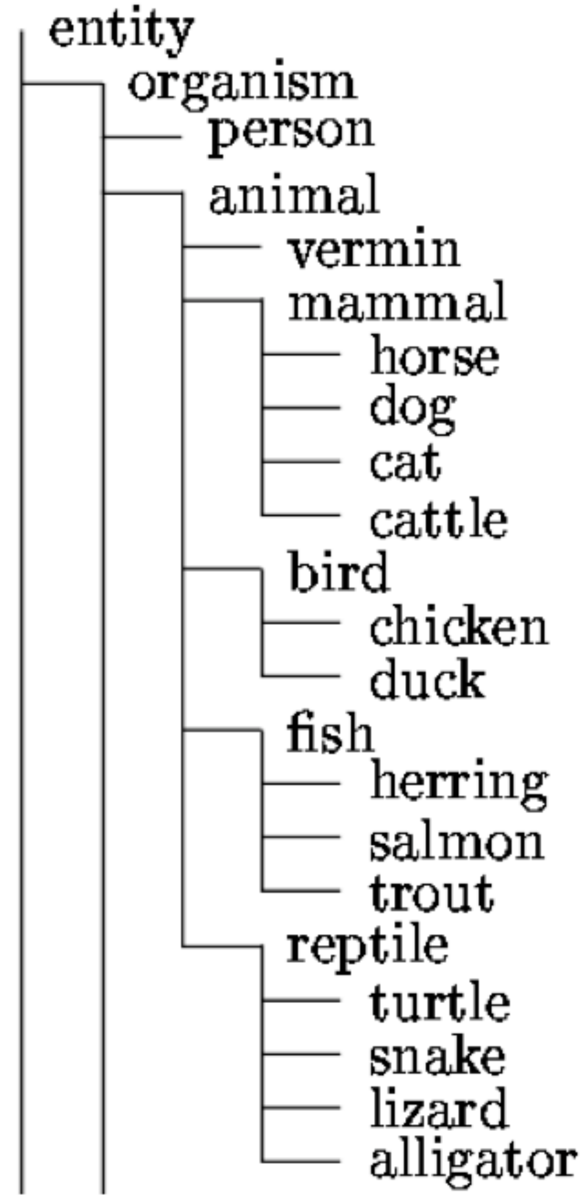
$Y \underset{H}{>} X$ if "X is a kind of Y".

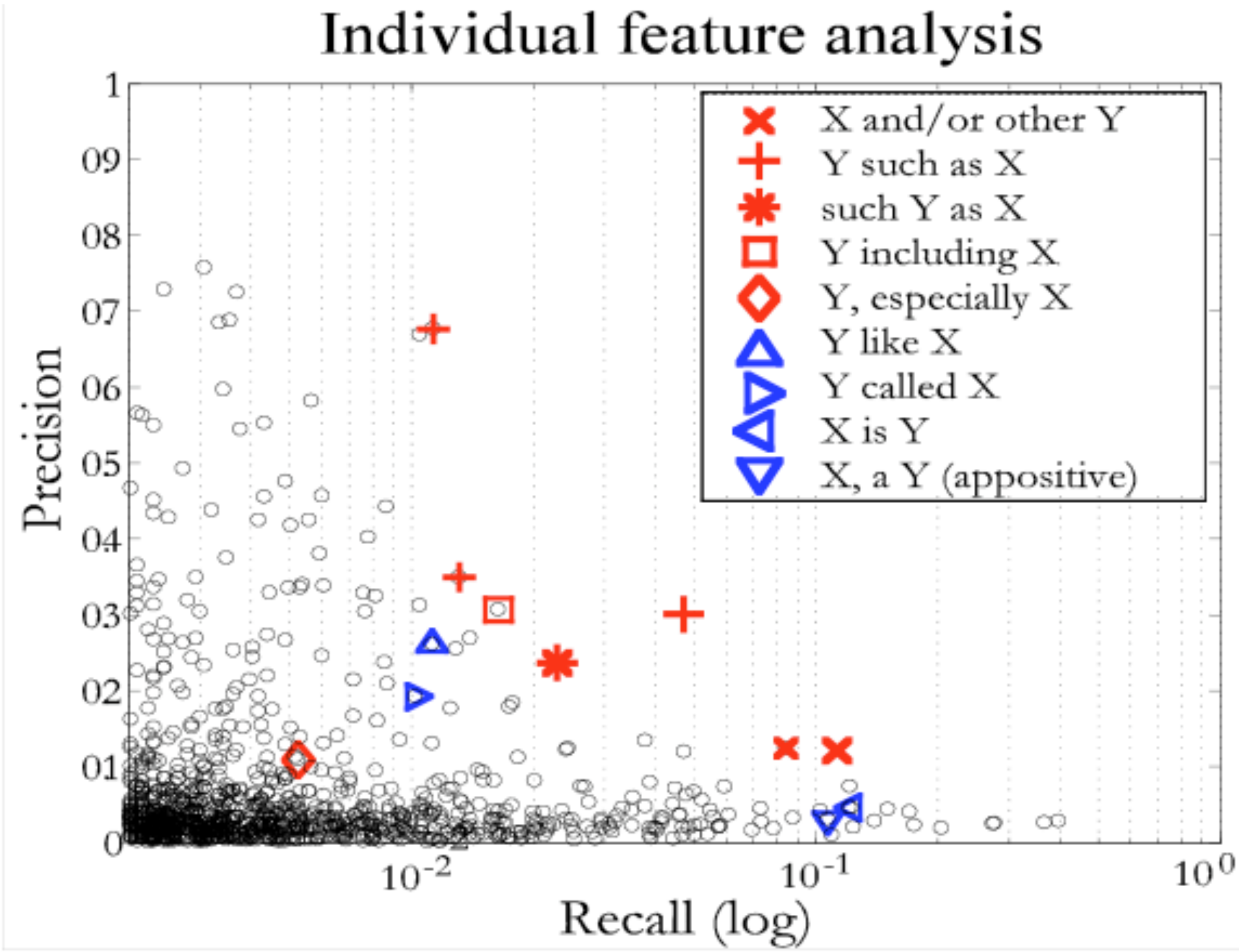
$$entity \underset{H}{>} organism \underset{H}{>} person$$

Coordinate Terms (taxonomic sisters)

$Y \underset{C}{\sim} X$ if X and Y possess a common hypernym, i.e. $\exists Z$ such that "X and Y are both kinds of Z."

$$horse \underset{C}{\sim} dog \underset{C}{\sim} cat$$

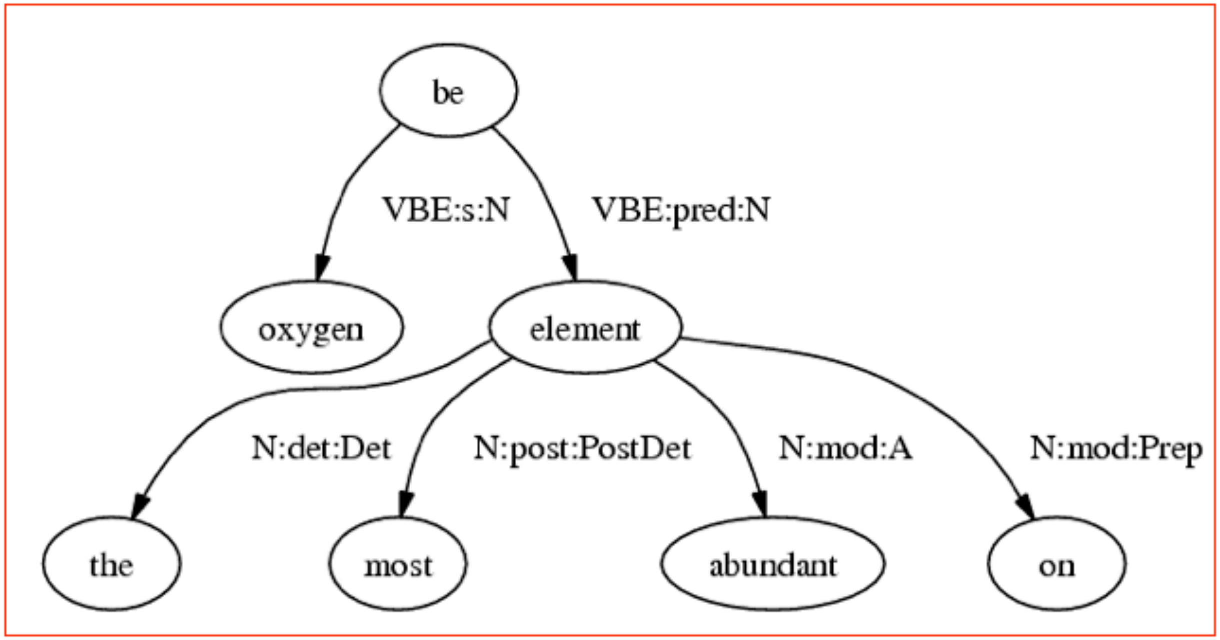




- Precision/recall for 69,592 classifiers (one per feature)
- Classifier *f* classifies noun pair ***x*** as hypernym iff $x_f > 0$
- **In red:** patterns originally proposed in (Hearst, 1992)

“Oxygen is the most abundant element on the moon.”

Dependency Graph:

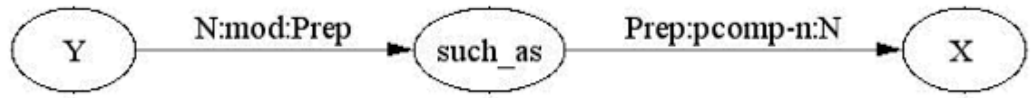


Dependency Paths (for “oxygen / element”):

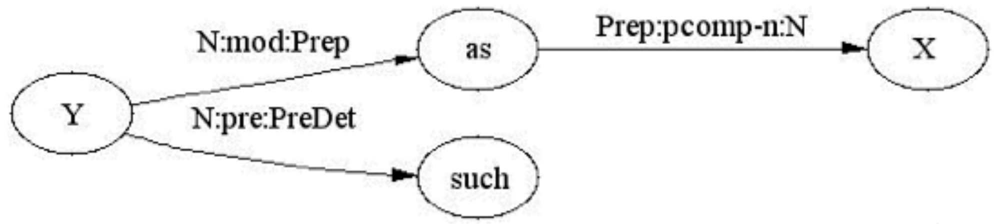
-N:s:VBE, “be” VBE:pred:N
-N:s:VBE, “be” VBE:pred:N,(the,Det:det:N)
-N:s:VBE, “be” VBE:pred:N,(most,PostDet:post:N)
-N:s:VBE, “be” VBE:pred:N,(abundant,A:mod:N)
-N:s:VBE, “be” VBE:pred:N,(on,Prep:mod:N)

Rediscovering Hearst's Patterns

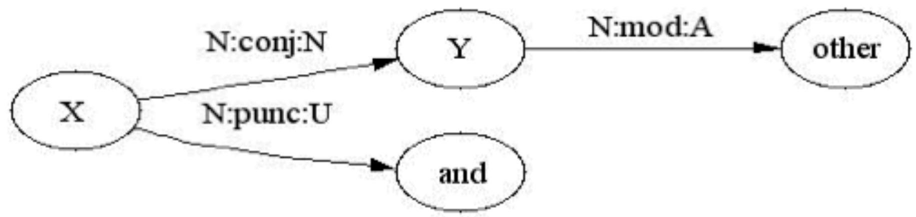
Y such as X…



Such Y as X…



X… and other Y



Proposed in (Hearst, 1992) and used in (Caraballo, 2001), (Widdows, 2003), and others – but what about the rest of the lexico-syntactic pattern space?

Example: Using the "Y called X" Pattern for Hypernym Acquisition

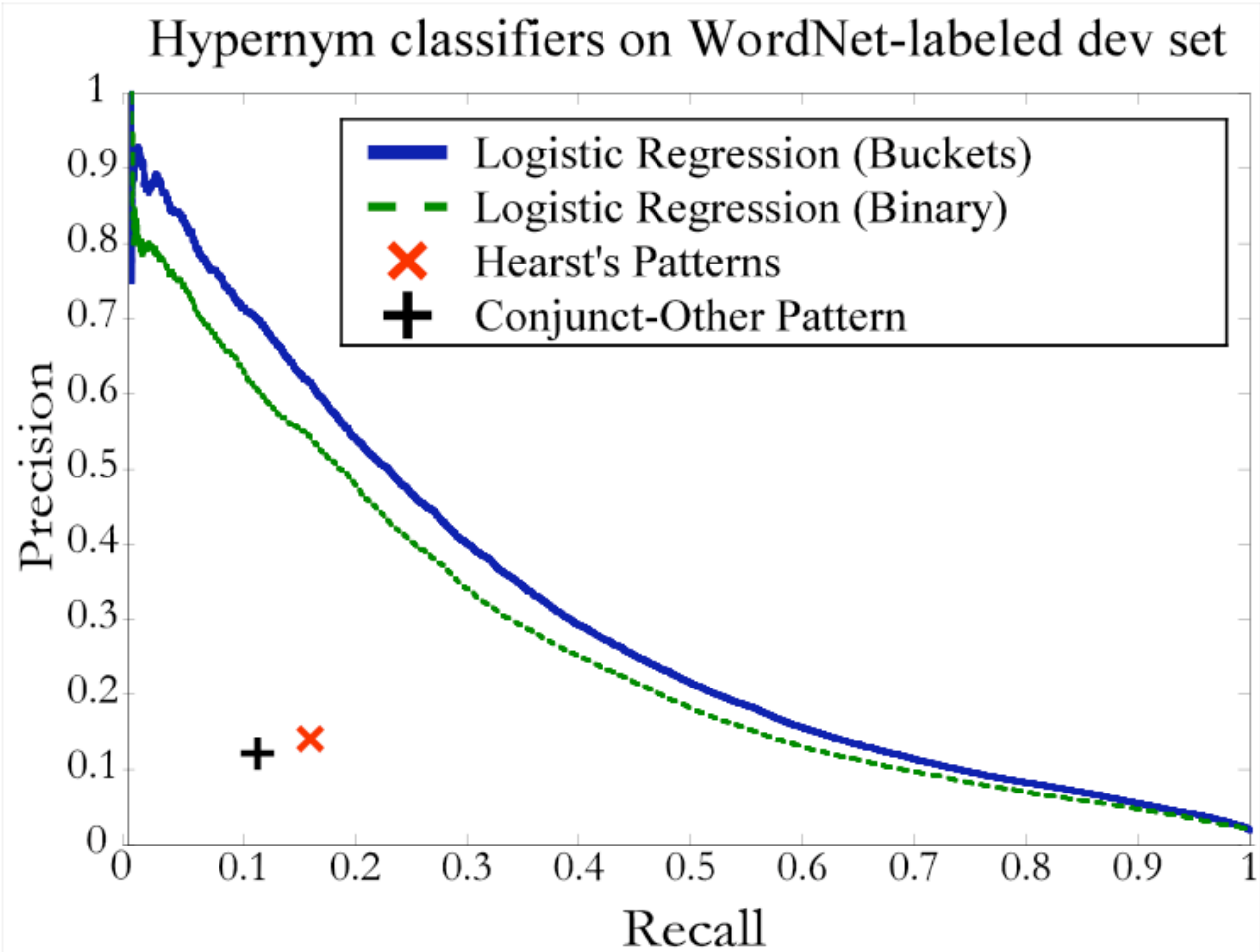
MINIPAR path: -N:desc:V,call,call,-V:vrel:N → "<hypernym> 'called' <hyponym>"

None of the following links are contained in WordNet (or the training set, by extension).

Hyponym	Hypernym	Sentence Fragment
efflorescence	**condition**	…and a **condition called efflorescence**…
'neal_inc	**company**	…The **company**, now **called O'Neal Inc.**…
hat_creek_outfit	**ranch**	…run a small **ranch called** the **Hat Creek Outfit**.
tardive_dyskinesia	**problem**	... irreversible **problem called tardive dyskinesia**…
hiv-1	**aids_virus**	…infected by the **AIDS virus**, **called HIV-1**.
bateau_mouche	**attraction**	…sightseeing **attraction called** the **Bateau Mouche**...
kibbutz_malkiyya	**collective_farm**	…Israeli **collective farm called Kibbutz Malkiyya**…

Type of Noun Pair	Count	Example Pair
NE: Person	7	"John F. Kennedy / president", "Marlin Fitzwater / spokesman"
NE: Place	7	"Diamond Bar / city", "France / place"
NE: Company	2	"American Can / company", "Simmons / company"
NE: Other	1	"Is Elvis Alive / book"
Not Named Entity:	9	"earthquake / disaster", "soybean / crop"

A better hypernym classifier



- 10-fold cross validation on the WordNet-labeled data
- **Conclusion:** 70,000 features are more powerful than 6

VERBOCEAN: Mining the Web for Fine-Grained Semantic Verb Relations

Timothy Chklovski and Patrick Pantel

Why Detect Semantic Rels between Verbs?

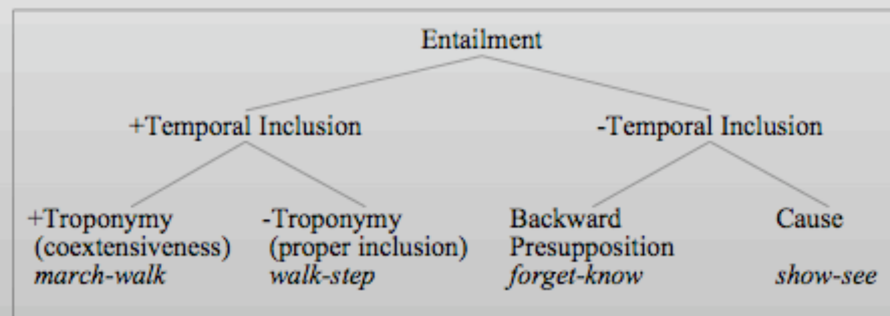
- So that we can
  - Understand the relationship when it's not stated
    - Napoleon ***fought*** and ***won*** the battle
    - During the holidays, people ***wrap*** and ***unwrap*** presents
    - Soldiers prefer to avoid getting ***wounded*** and ***killed***
  - Use the relationship when summarizing across documents (e.g. same event, preceding event)
    - The board ***considered*** the offer of $3B
    - The board ***accepted*** the offer $3.8B
    - The board ***okayed*** the offer of approximately $4B
  - Determine if two people have similar views on and event
    - "I ***nudged*** him."
    - "He ***shoved*** me."
- Hard to do manually

Why use Web? Motivating Intuition

- Small collections are tough: Semantics is often implied (Lenat, Chklovski)
- The Web's **10^{12}** is a lot of words
- So, Use small bits of more detailed text to help with mass of general text
  - Patterns issued to a search engine and their correlation

Relevant Work

- Levin's classes (similarity)
  - 3200 verbs in 191 classes
- PropBank
  - 4,659 framesets (1.4 framesets per verb)
- VerbNet
  - 191 coarse-grained groupings (with overlap)
- FrameNet
- WordNet
  - troponomy
  - antonymy
  - entailment
  - cause



Fellbaum's (1998) entailment hierarchy.

VerbOcean: Web-based Extraction of Verb Relations

- VerbOcean is a network of verb relations
  - Currently, over 3400 nodes with on average 13 relations per verb
- Detected relation types are:
  - similarity
  - strength
  - antonymy
  - enablement
  - temporal precedence (happens-before)
- Download from http://semantics.isi.edu/ocean/

Approach

- Three stages:
  - Identify pairs of highly associated verbs co-occurring on the Web with sufficient frequency using DIRT (Lin and Pantel 2001)
  - For each verb pair
    - test patterns associated with each semantic relation
      - E.g. Temporal Precedence:
        "to *X* and then *Y*", "*X*ed and then *Y*ed"
    - calculate a score for each possible semantic relation
  - Compare the strengths of the individual semantic relations and output a consistent set as the final output
    - prefer the most specific and then strongest relations

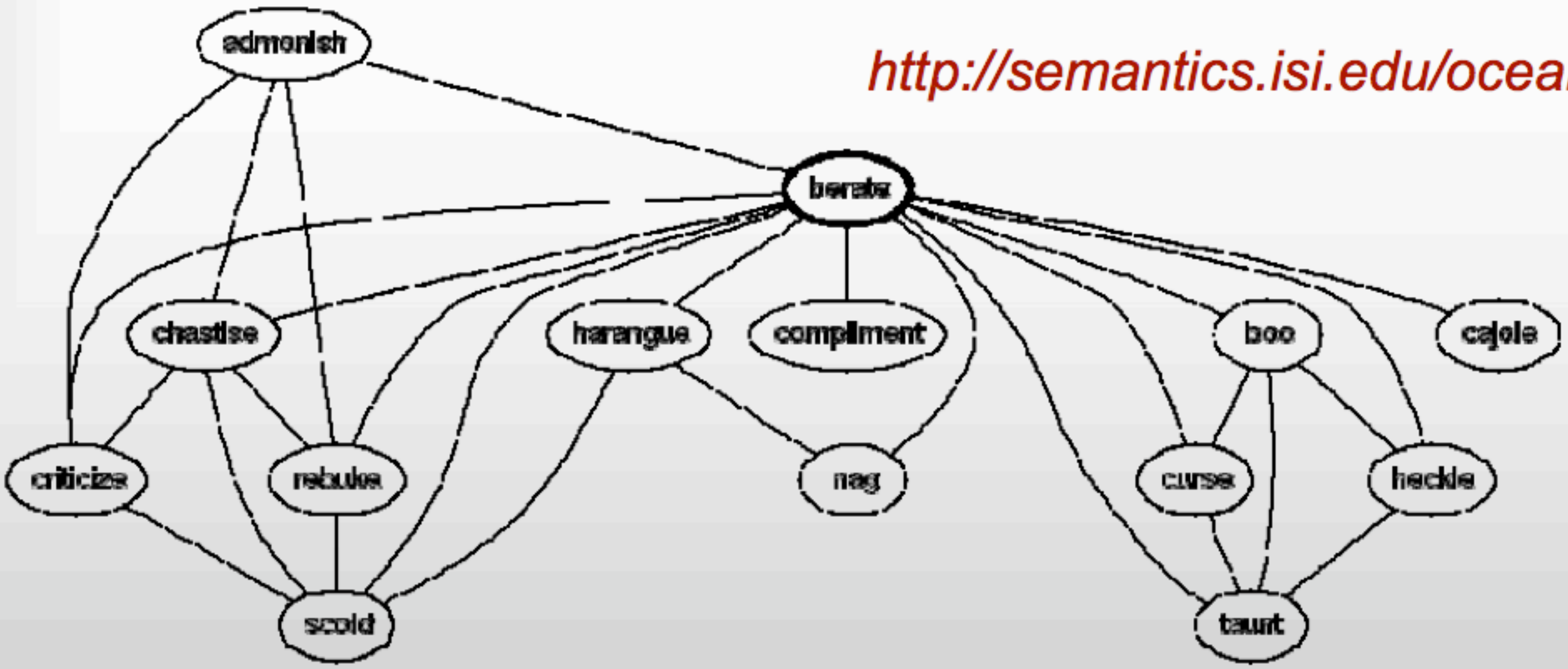
Lexical Patterns

SEMANTIC RELATION	*Surface Patterns*	*Example*
similarity (4)	X ie Y Xed and Yed	*“She **heckled and taunted** the comedian.”*
strength (8)	X even Y Xed even Yed Xed and even Yed not just Xed but Yed	*“He not **just harassed, but terrorized** her.”*
enablement (4)	Xed * by Ying the Xed * by Ying or to X * by Ying the	*“She **saved the document by clicking the** button.”*
antonymy (7)	either X or Y either Xs or Ys Xed * but Yed	*“There’s something about Mary: you will **either love or hate** her.”*
happens-before (12)	to X and then Y Xed * and then Yed to X and later Y to X and subsequently Y Xed and subsequently Yed	*“He **designed the prototype and then patented** it.”*

Lexical Patterns Match...

- Refined to decrease capturing wrong parts of speech or incorrect semantic relations
  - Xed * by Ying the; Xed * by Ying or
    - "… waved at by parking guard …"
    - "… encouraged further by sailing lessons …"

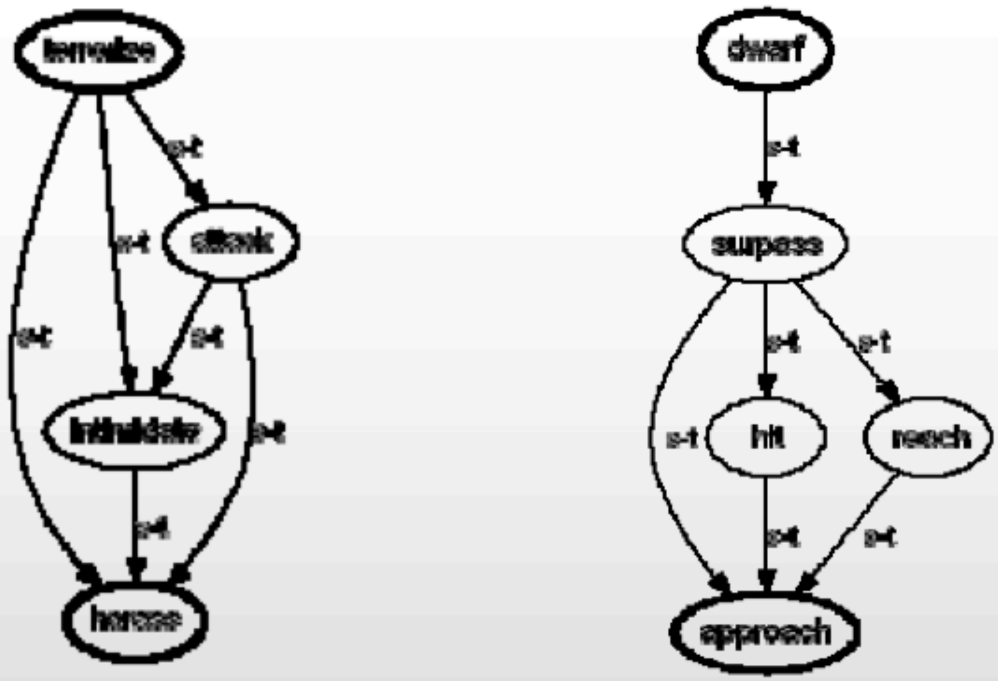
VerbOcean – Similarity



http://semantics.isi.edu/ocean/

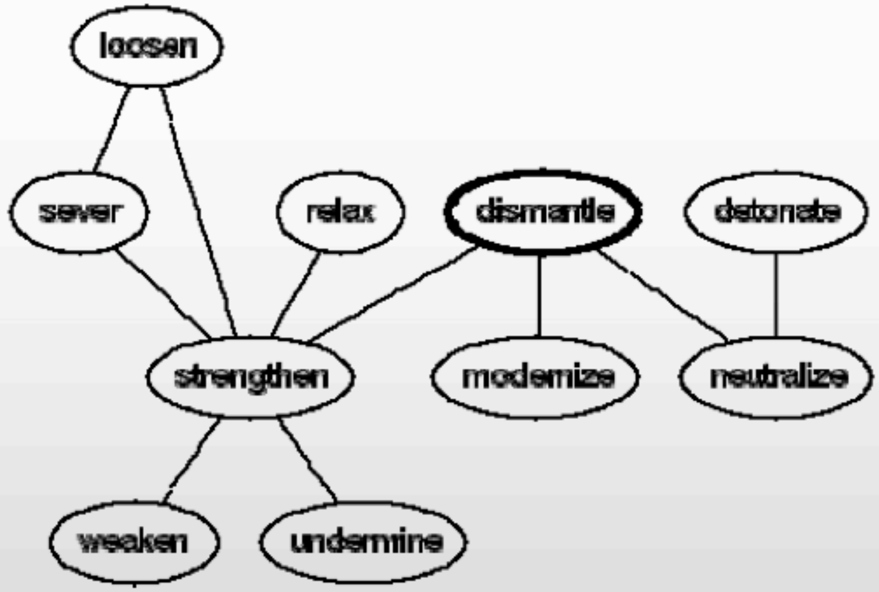
- Verbs that are similar or related
  - e.g. boo - heckle

VerbOcean – Strength



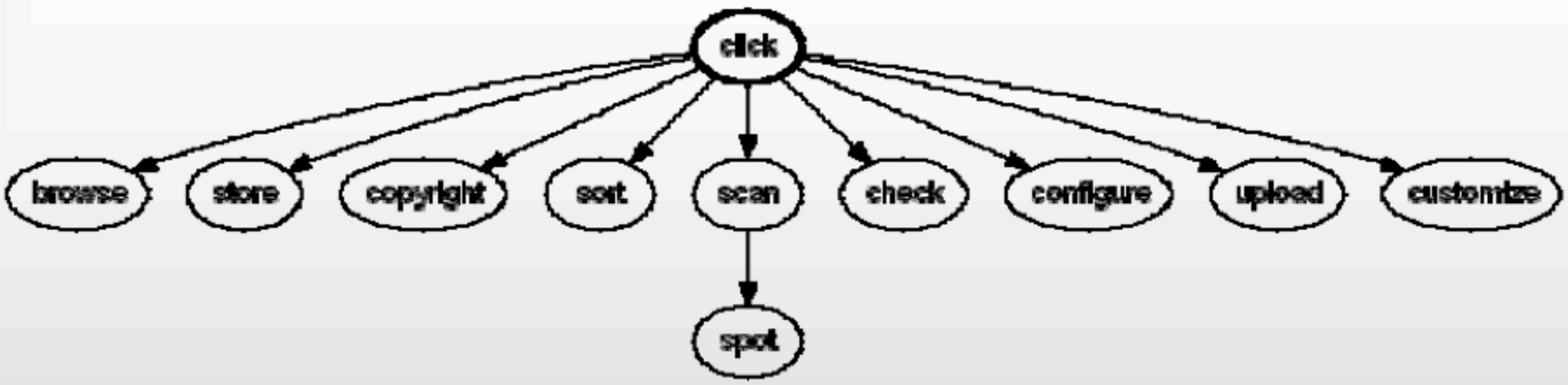
- Similar verbs that denote a more intense, thorough, comprehensive or absolute action
  - e.g. change-of-state verbs that denote a more complete change (shock → startle)

VerbOcean – Antonymy



- Semantic opposition
  - switching thematic roles associated with the verb (buy – sell)
  - stative verbs (live – die)
  - sibling verbs which share a parent (walk – run)
  - restitutive opposition: antonymy + happens-before (damage - repair)

VerbOcean – Enablement



- Holds between two verbs V_1 and V_2 when the pair can be glossed as "V_1 is accomplished by V_2" (assess - review)

Appendix. Sample relations extracted by our system.

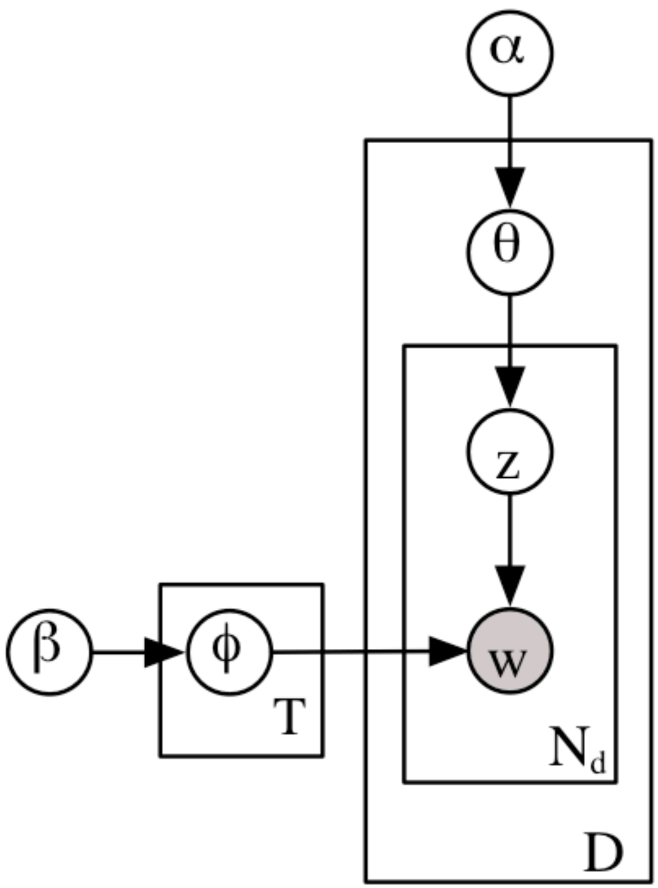
SEMANTIC RELATION	EXAMPLES	SEMANTIC RELATION	EXAMPLES	SEMANTIC RELATION	EXAMPLES
similarity	maximize :: enhance produce :: create reduce :: restrict	enablement	assess :: review accomplish :: complete double-click :: click	happens before	detain :: prosecute enroll :: graduate schedule :: reschedule
strength	permit :: authorize surprise :: startle startle :: shock	antonymy	assemble :: dismantle regard :: condemn roast :: fry		

Demo

http://semantics.isi.edu/ocean/

Clustering words into topics with Latent Dirichlet Allocation

[Blei, Ng, Jordan 2003]



Generative Process:

For each document:

Sample a distribution over topics, θ

For each word in doc

Sample a topic, z

Sample a word from the topic, w

Example:

70% Iraq war
30% US election

Iraq war

"bombing"

Example topics induced from a large collection of text

DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

Example topics
induced from a large collection of text

DISEASE	WATER	MIND	STORY	**FIELD**	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	**FIELD**	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	**FIELD**	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	**FIELD**
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

Collocations

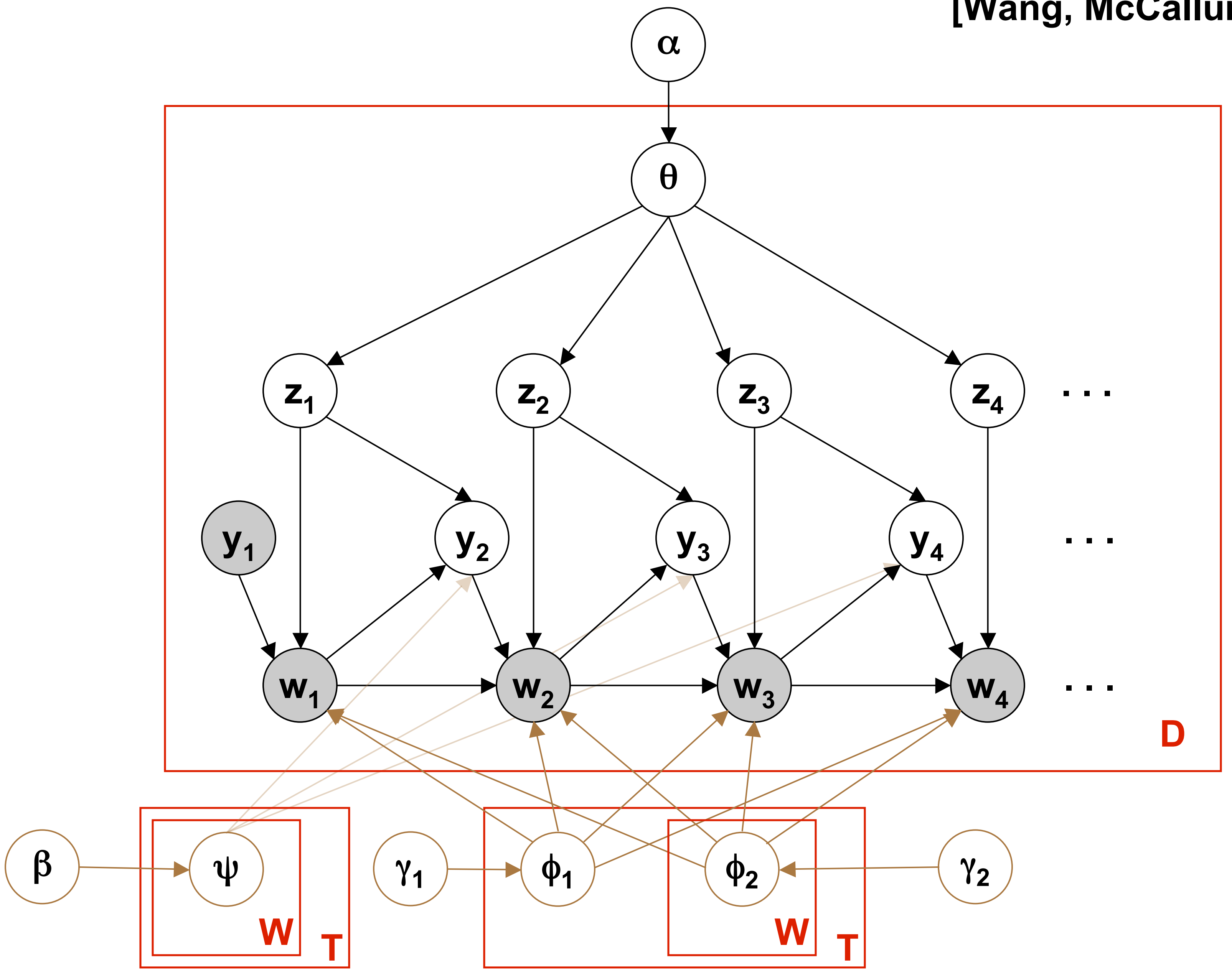
- An expression consisting of two or more words that correspond to some conventional way of saying things.
- Characterized by limited *compositionality*.
  - *compositional*: meaning of expression can be predicted by meaning of its parts.
  - “strong tea”, “rich in calcium”
  - “weapons of mass destruction”
  - “kick the bucket”, “hear it through the grapevine”

Topics Modeling Phrases

- Topics based only on unigrams often difficult to interpret
- Topic discovery itself is confused because important meaning / distinctions carried by phrases.
- Significant opportunity to provide improved language models to ASR, MT, IR, etc.

Topical N-gram Model

[Wang, McCallum 2005]



LDA Topic

LDA	Topical N-grams
algorithms	genetic algorithms
algorithm	genetic algorithm
genetic	evolutionary computation
problems	evolutionary algorithms
efficient	fitness function

Topic Comparison

LDA	Topical N-grams (2)	Topical N-grams (1)
learning	reinforcement learning	policy
optimal	optimal policy	action
reinforcement	dynamic programming	states
state	optimal control	actions
problems	function approximator	function
policy	prioritized sweeping	reward
dynamic	finite-state controller	control
action	learning system	agent
programming	reinforcement learning rl	q-learning
actions	function approximators	optimal
function	markov decision problems	goal
markov	markov decision processes	learning
methods	local search	space
decision	state-action pair	step
rl	markov decision process	environment
continuous	belief states	system
spaces	stochastic policy	problem
step	action selection	steps
policies	upright position	sutton
planning	reinforcement learning methods	policies

Topic Comparison

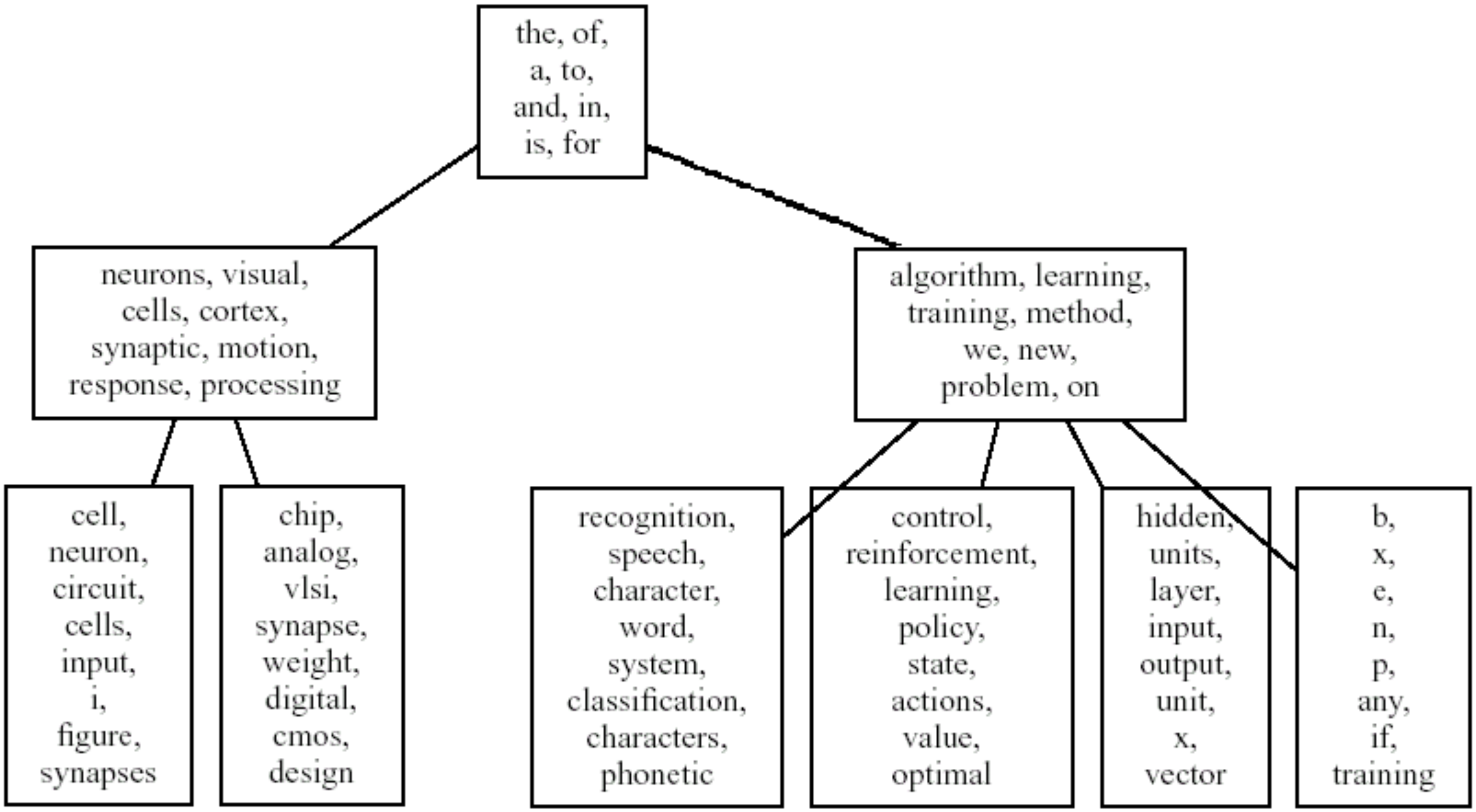
LDA	Topical N-grams (2)	Topical N-grams (1)
motion	receptive field	motion
visual	spatial frequency	response
field	temporal frequency	direction
position	visual motion	cells
figure	motion energy	stimulus
direction	tuning curves	figure
fields	horizontal cells	contrast
eye	motion detection	velocity
location	preferred direction	model
retina	visual processing	responses
receptive	area mt	stimuli
velocity	visual cortex	moving
vision	light intensity	cell
moving	directional selectivity	intensity
system	high contrast	population
flow	motion detectors	image
edge	spatial phase	center
center	moving stimuli	tuning
light	decision strategy	complex
local	visual stimuli	directions

Topic Comparison

LDA	Topical N-grams (2)	Topical N-grams (1)
word	speech recognition	speech
system	training data	word
recognition	neural network	training
hmm	error rates	system
speech	neural net	recognition
training	hidden markov model	hmm
performance	feature vectors	speaker
phoneme	continuous speech	performance
words	training procedure	phoneme
context	continuous speech recognition	acoustic
systems	gamma filter	words
frame	hidden control	context
trained	speech production	systems
speaker	neural nets	frame
sequence	input representation	trained
speakers	output layers	sequence
mlp	training algorithm	phonetic
frames	test set	speakers
segmentation	speech frames	mlp
models	speaker dependent	hybrid

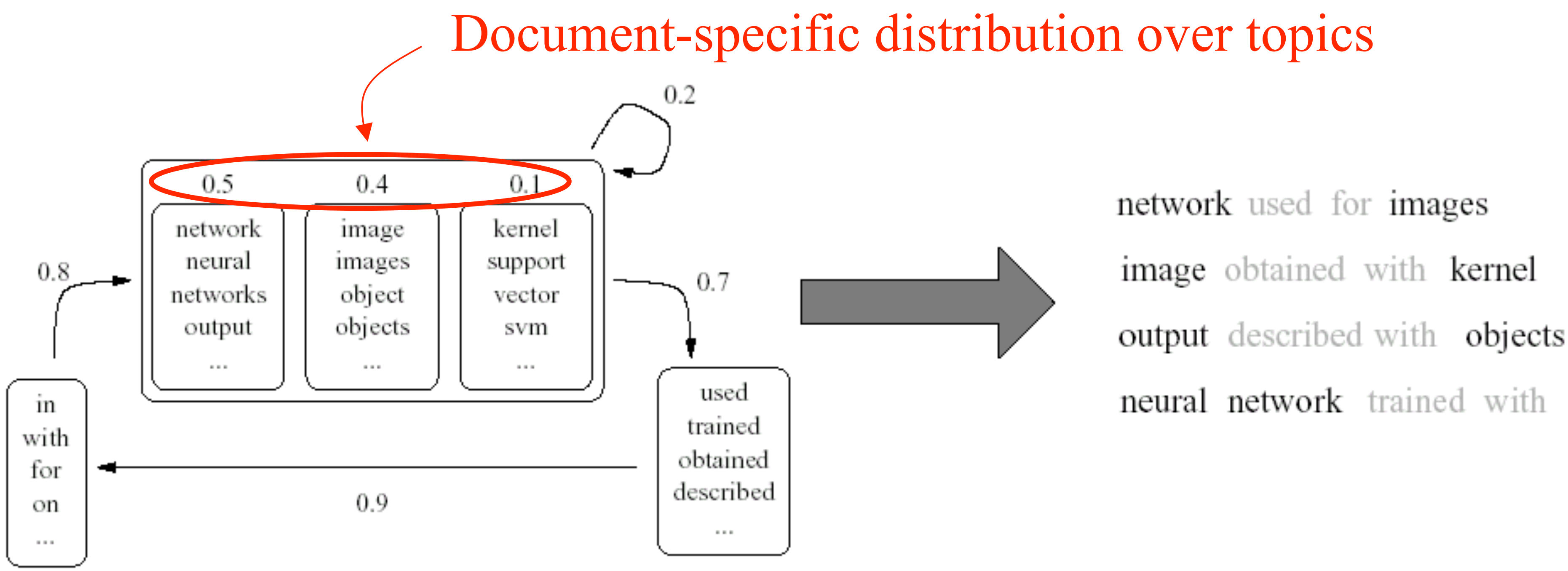
Unsupervised learning of topic hierarchies

(Blei, Griffiths, Jordan & Tenenbaum, NIPS 2003)

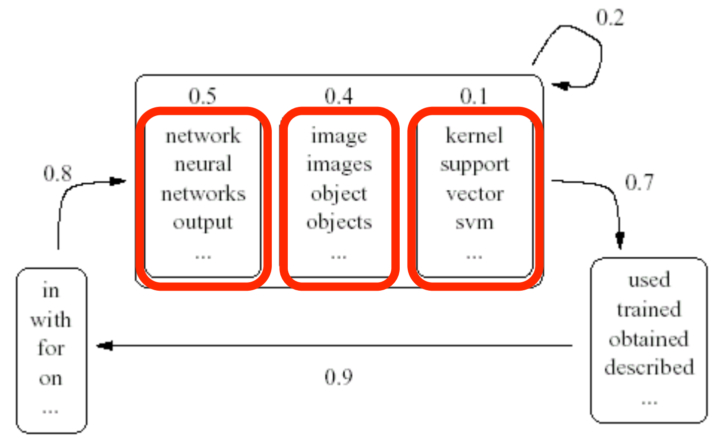


Joint models of syntax and semantics (Griffiths, Steyvers, Blei & Tenenbaum, NIPS 2004)

- Embed topics model inside an *n*th order Hidden Markov Model:

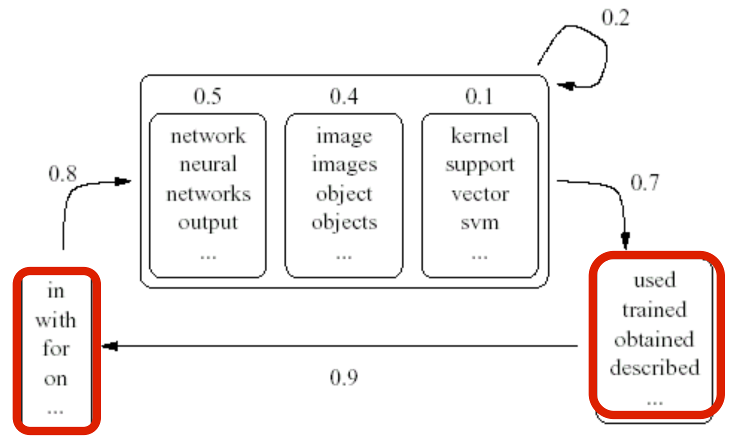


Semantic classes



FOOD	MAP	DOCTOR	BOOK	GOLD	BEHAVIOR	CELLS	PLANTS
FOODS	NORTH	PATIENT	BOOKS	IRON	SELF	CELL	PLANT
BODY	EARTH	HEALTH	READING	SILVER	INDIVIDUAL	ORGANISMS	LEAVES
NUTRIENTS	SOUTH	HOSPITAL	INFORMATION	COPPER	PERSONALITY	ALGAE	SEEDS
DIET	POLE	MEDICAL	LIBRARY	METAL	RESPONSE	BACTERIA	SOIL
FAT	MAPS	CARE	REPORT	METALS	SOCIAL	MICROSCOPE	ROOTS
SUGAR	EQUATOR	PATIENTS	PAGE	STEEL	EMOTIONAL	MEMBRANE	FLOWERS
ENERGY	WEST	NURSE	TITLE	CLAY	LEARNING	ORGANISM	WATER
MILK	LINES	DOCTORS	SUBJECT	LEAD	FEELINGS	FOOD	FOOD
EATING	EAST	MEDICINE	PAGES	ADAM	PSYCHOLOGISTS	LIVING	GREEN
FRUITS	AUSTRALIA	NURSING	GUIDE	ORE	INDIVIDUALS	FUNGI	SEED
VEGETABLES	GLOBE	TREATMENT	WORDS	ALUMINUM	PSYCHOLOGICAL	MOLD	STEMS
WEIGHT	POLES	NURSES	MATERIAL	MINERAL	EXPERIENCES	MATERIALS	FLOWER
FATS	HEMISPHERE	PHYSICIAN	ARTICLE	MINE	ENVIRONMENT	NUCLEUS	STEM
NEEDS	LATITUDE	HOSPITALS	ARTICLES	STONE	HUMAN	CELLED	LEAF
CARBOHYDRATES	PLACES	DR	WORD	MINERALS	RESPONSES	STRUCTURES	ANIMALS
VITAMINS	LAND	SICK	FACTS	POT	BEHAVIORS	MATERIAL	ROOT
CALORIES	WORLD	ASSISTANT	AUTHOR	MINING	ATTITUDES	STRUCTURE	POLLEN
PROTEIN	COMPASS	EMERGENCY	REFERENCE	MINERS	PSYCHOLOGY	GREEN	GROWING
MINERALS	CONTINENTS	PRACTICE	NOTE	TIN	PERSON	MOLDS	GROW

Syntactic classes



SAID	THE	MORE	ON	GOOD	ONE	HE	BE
ASKED	HIS	SUCH	AT	SMALL	SOME	YOU	MAKE
THOUGHT	THEIR	LESS	INTO	NEW	MANY	THEY	GET
TOLD	YOUR	MUCH	FROM	IMPORTANT	TWO	I	HAVE
SAYS	HER	KNOWN	WITH	GREAT	EACH	SHE	GO
MEANS	ITS	JUST	THROUGH	LITTLE	ALL	WE	TAKE
CALLED	MY	BETTER	OVER	LARGE	MOST	IT	DO
CRIED	OUR	RATHER	AROUND	*	ANY	PEOPLE	FIND
SHOWS	THIS	GREATER	AGAINST	BIG	THREE	EVERYONE	USE
ANSWERED	THESE	HIGHER	ACROSS	LONG	THIS	OTHERS	SEE
TELLS	A	LARGER	UPON	HIGH	EVERY	SCIENTISTS	HELP
REPLIED	AN	LONGER	TOWARD	DIFFERENT	SEVERAL	SOMEONE	KEEP
SHOUTED	THAT	FASTER	UNDER	SPECIAL	FOUR	WHO	GIVE
EXPLAINED	NEW	EXACTLY	ALONG	OLD	FIVE	NOBODY	LOOK
LAUGHED	THOSE	SMALLER	NEAR	STRONG	BOTH	ONE	COME
MEANT	EACH	SOMETHING	BEHIND	YOUNG	TEN	SOMETHING	WORK
WROTE	MR	BIGGER	OFF	COMMON	SIX	ANYONE	MOVE
SHOWED	ANY	FEWER	ABOVE	WHITE	MUCH	EVERYBODY	LIVE
BELIEVED	MRS	LOWER	DOWN	SINGLE	TWENTY	SOME	EAT
WHISPERED	ALL	ALMOST	BEFORE	CERTAIN	EIGHT	THEN	BECOME

Corpus-specific factorization (NIPS)

Semantics	image	data	state	membrane	chip	experts	kernel	network
	images	gaussian	policy	synaptic	analog	expert	support	neural
	object	mixture	value	cell	neuron	gating	vector	networks
	objects	likelihood	function	*	digital	hme	svm	output
	feature	posterior	action	current	synapse	architecture	kernels	input
	recognition	prior	reinforcement	dendritic	neural	mixture	#	training
	views	distribution	learning	potential	hardware	learning	space	inputs
	#	em	classes	neuron	weight	mixtures	function	weights
	pixel	bayesian	optimal	conductance	#	function	machines	#
	visual	parameters	*	channels	vlsi	gate	set	outputs
Syntax	in	is	see	used	model	networks	however	#
	with	was	show	trained	algorithm	values	also	*
	for	has	note	obtained	system	results	then	i
	on	becomes	consider	described	case	models	thus	x
	from	denotes	assume	given	problem	parameters	therefore	t
	at	being	present	found	network	units	first	n
	using	remains	need	presented	method	data	here	-
	into	represents	propose	defined	approach	functions	now	c
	over	exists	describe	generated	paper	problems	hence	r
	within	seems	suggest	shown	process	algorithms	finally	p

Syntactic classes in PNAS

5	8	14	25	26	30	33
IN	ARE	THE	SUGGEST	LEVELS	RESULTS	BEEN
FOR	WERE	THIS	INDICATE	NUMBER	ANALYSIS	MAY
ON	WAS	ITS	SUGGESTING	LEVEL	DATA	CAN
BETWEEN	IS	THEIR	SUGGESTS	RATE	STUDIES	COULD
DURING	WHEN	AN	SHOWED	TIME	STUDY	WELL
AMONG	REMAIN	EACH	REVEALED	CONCENTRATIONS	FINDINGS	DID
FROM	REMAINS	ONE	SHOW	VARIETY	EXPERIMENTS	DOES
UNDER	REMAINED	ANY	DEMONSTRATE	RANGE	OBSERVATIONS	DO
WITHIN	PREVIOUSLY	INCREASED	INDICATING	CONCENTRATION	HYPOTHESIS	MIGHT
THROUGHOUT	BECOME	EXOGENOUS	PROVIDE	DOSE	ANALYSES	SHOULD
THROUGH	BECAME	OUR	SUPPORT	FAMILY	ASSAYS	WILL
TOWARD	BEING	RECOMBINANT	INDICATES	SET	POSSIBILITY	WOULD
INTO	BUT	ENDOGENOUS	PROVIDES	FREQUENCY	MICROSCOPY	MUST
AT	GIVE	TOTAL	INDICATED	SERIES	PAPER	CANNOT
INVOLVING	MERE	PURIFIED	DEMONSTRATED	AMOUNTS	WORK	REMAINED
AFTER	APPEARED	TILE	SHOWS	RATES	EVIDENCE	ALSO
ACROSS	APPEAR	FULL	SO	CLASS	FINDING	THEY
AGAINST	ALLOWED	CHRONIC	REVEAL	VALUES	MUTAGENESIS	BECOME
WHEN	NORMALLY	ANOTHER	DEMONSTRATES	AMOUNT	OBSERVATION	MAG
ALONG	EACH	EXCESS	SUGGESTED	SITES	MEASUREMENTS	LIKELY

Semantic highlighting

Darker words are more likely to have been generated from the topic-based "semantics" module:

In contrast to this approach, we study here how the overall network activity can control single cell parameters such as input resistance, as well as time and space constants, parameters that are crucial for excitability and spariotemporal (sic) integration.

The integrated architecture in this paper combines feed forward control and error feedback adaptive control using neural networks.

In other words, for our proof of convergence, we require the softassign algorithm to return a doubly stochastic matrix as *sinkhorn theorem guarantees that it will instead of a matrix which is merely close to being doubly stochastic based on some reasonable metric.

The aim is to construct a portfolio with a maximal expected return for a given risk level and time horizon while simultaneously obeying *institutional or *legally required constraints.

The left graph is the standard experiment the right from a training with # samples.

The graph G is called the *guest graph, and H is called the host graph.

Social Network Analysis: Group and Topic Discovery

Xuerui Wang and Andrew McCallum

Groups and Topics

- Input:
  - Observed relations between people
  - Attributes on those relations (text, or categorical)
- Output:
  - Attributes clustered into “topics”
  - Groups of people---varying depending on topic

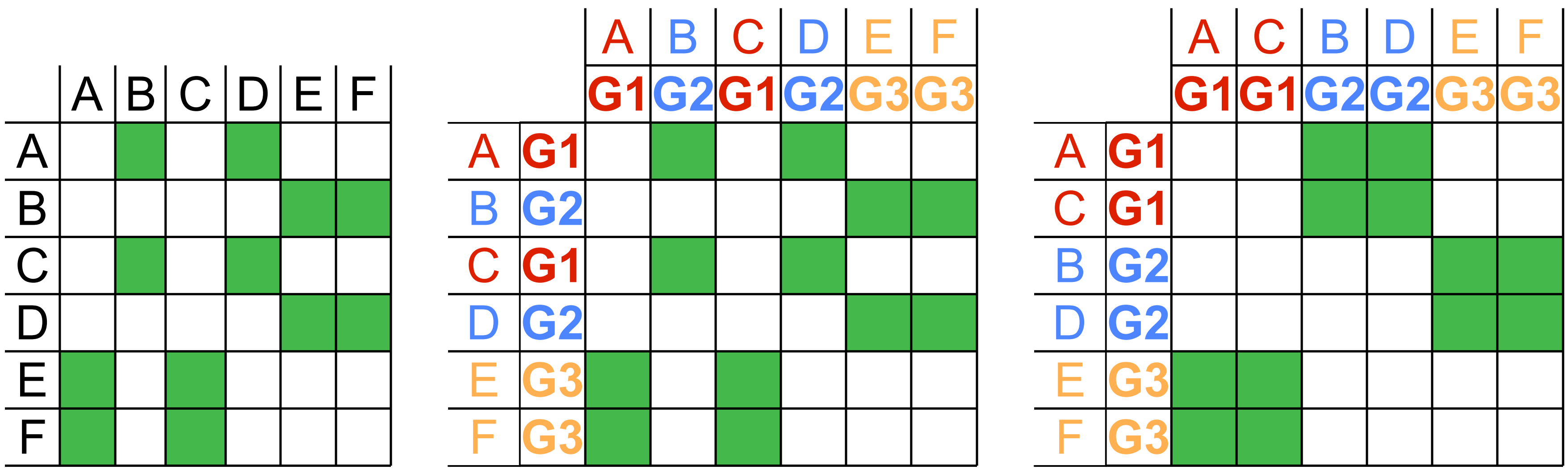
Discovering Groups from Observed Set of Relations

Student Roster	**Academic Admiration**
Adams **B**ennett **C**arter **D**avis **E**dwards **F**rederking	Acad(A, B) Acad(C, B) Acad(A, D) Acad(C, D) Acad(B, E) Acad(D, E) Acad(B, F) Acad(D, F) Acad(E, A) Acad(F, A) Acad(E, C) Acad(F, C)

Admiration relations among six high school students.

Adjacency Matrix Representing Relations

Student Roster	**Academic Admiration**
Adams **B**ennett **C**arter **D**avis **E**dwards **F**rederking	Acad(A, B) Acad(C, B) Acad(A, D) Acad(C, D) Acad(B, E) Acad(D, E) Acad(B, F) Acad(D, F) Acad(E, A) Acad(F, A) Acad(E, C) Acad(F, C)

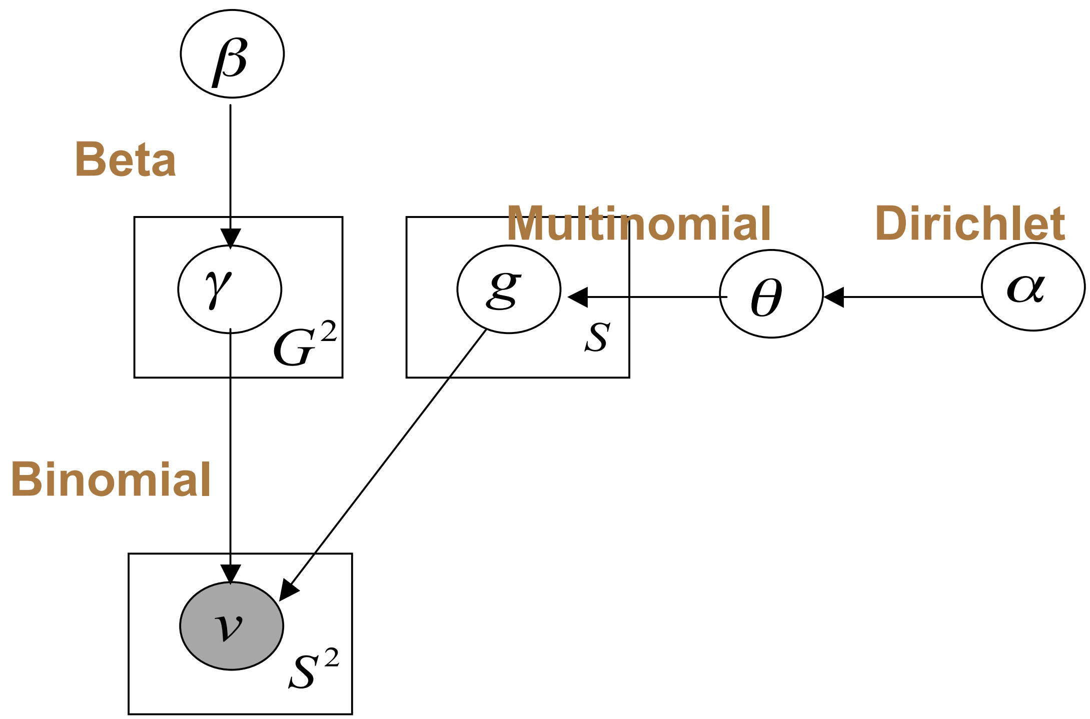


Group Model:
Partitioning Entities into Groups

Stochastic Blockstructures for Relations
[Nowicki, Snijders 2001]

S: number of entities

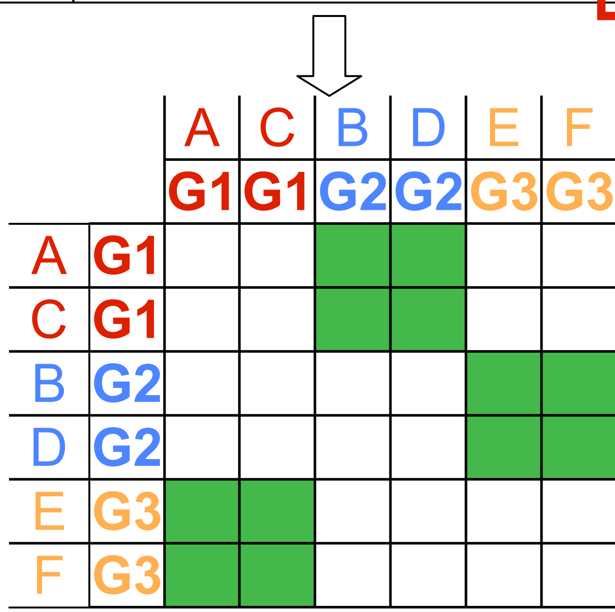
G: number of groups

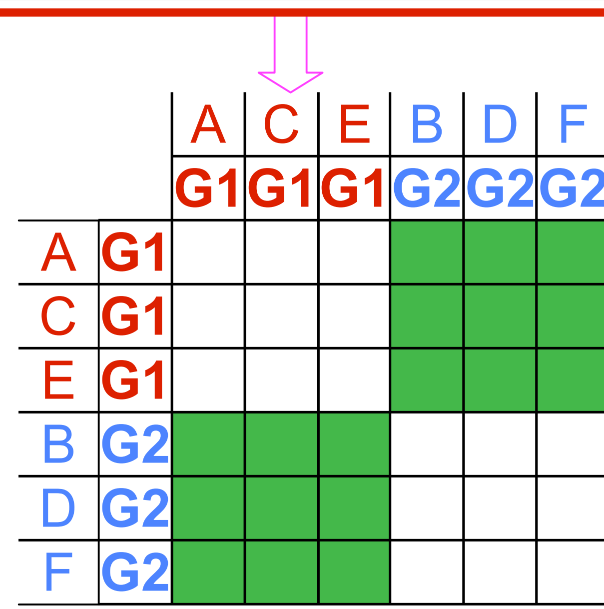


Enhanced with arbitrary number of groups in [Kemp, Griffiths, Tenenbaum 2004]

Two Relations with Different Attributes

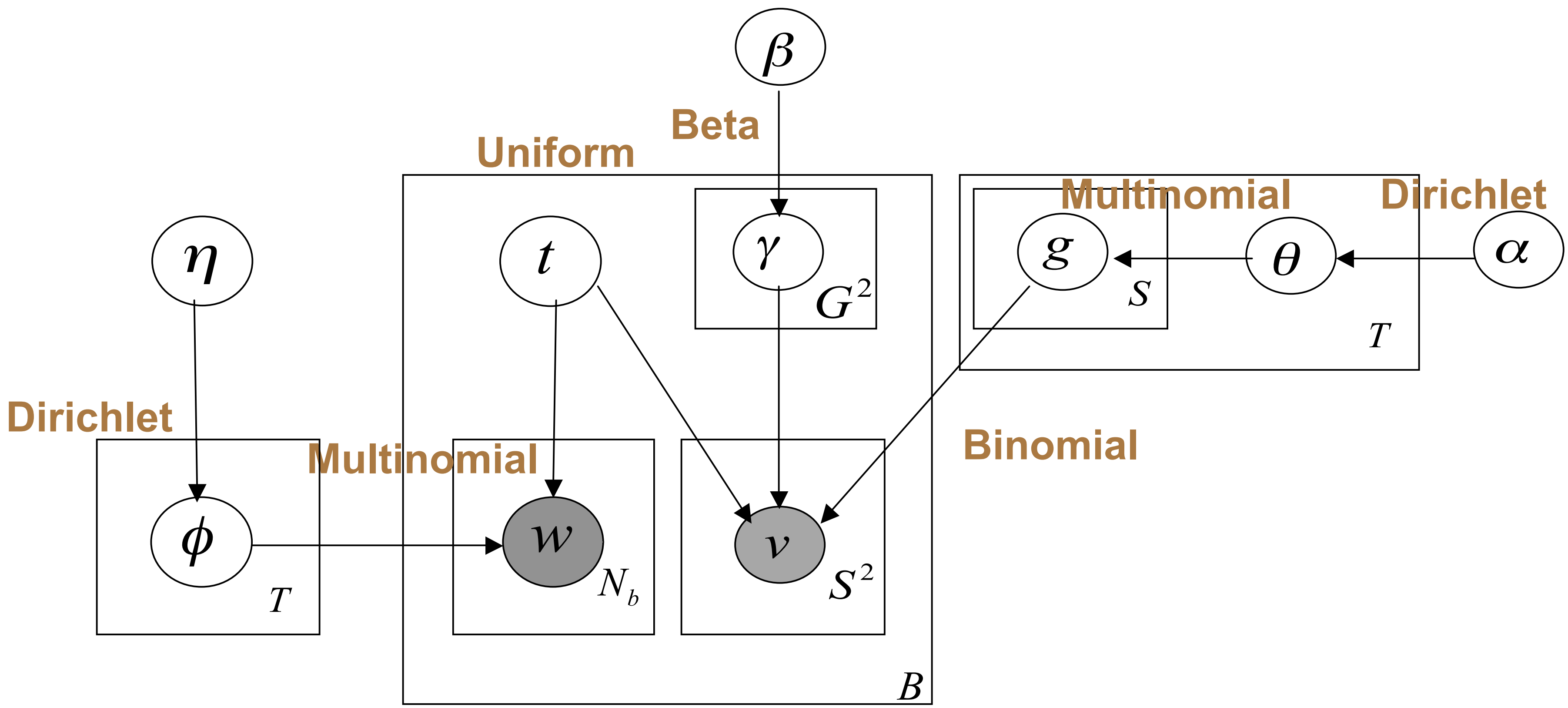
Student Roster	Academic Admiration	Social Admiration
Adams	Acad(A, B) Acad(C, B)	Soci(A, B) Soci(A, D) Soci(A, F)
Bennett	Acad(A, D) Acad(C, D)	Soci(B, A) Soci(B, C) Soci(B, E)
Carter	Acad(B, E) Acad(D, E)	Soci(C, B) Soci(C, D) Soci(C, F)
Davis	Acad(B, F) Acad(D, F)	Soci(D, A) Soci(D, C) Soci(D, E)
Edwards	Acad(E, A) Acad(F, A)	Soci(E, B) Soci(E, D) Soci(E, F)
Frederking	Acad(E, C) Acad(F, C)	Soci(F, A) Soci(F, C) Soci(F, E)





The Group-Topic Model: Discovering Groups and Topics Simultaneously

[Wang, Mohanty, McCallum 2006]



Inference and Estimation

Gibbs Sampling:

- Many r.v.s can be integrated out
- Easy to implement
- Reasonably fast

We assume the relationship is symmetric.

$$P(t_b|\mathbf{V},\mathbf{g},\mathbf{w},\mathbf{t}_{-b},\alpha,\beta,\eta)$$

$$\propto \frac{\prod_{v=1}^{V}\prod_{x=1}^{e_v^{(b)}}(\eta_v + c_{t_b v} - x)}{\prod_{x=1}^{\sum_{v=1}^{V} e_v^{(b)}}\left(\sum_{v=1}^{V}(\eta_v + c_{t_b v}) - x\right)} \times \prod_{g=1}^{G}\prod_{h=g}^{G}\frac{\prod_{k=1}^{2}\Gamma(\beta_k + m_{ghk}^{(b)})}{\Gamma(\sum_{k=1}^{2}(\beta_k + m_{ghk}^{(b)}))},$$

$$P(g_{st}|\mathbf{V},\mathbf{g}_{-st},\mathbf{w},\mathbf{t},\alpha,\beta,\eta)$$

$$\propto \frac{\alpha_{g_{st}} + n_{tg_{st}} - 1}{\sum_{g=1}^{G}(\alpha_g + n_{tg}) - 1}\prod_{b=1}^{B}\left(I(t_b = t) \times \prod_{h=1}^{G}\frac{\prod_{k=1}^{2}\prod_{x=1}^{d_{g_{st}hk}^{(b)}}\left(\beta_k + m_{g_{st}hk}^{(b)} - x\right)}{\prod_{x=1}^{\sum_{k=1}^{2} d_{g_{st}hk}^{(b)}}\left((\sum_{k=1}^{2}(\beta_k + m_{g_{st}hk}^{(b)}) - x\right)}\right)$$

Dataset #1:

U.S. Senate

- **16 years of voting records in the US Senate (1989 – 2005)**
- **a Senator may respond *Yea* or *Nay* to a resolution**
- **3423 resolutions with text attributes (index terms)**
- **191 Senators in total across 16 years**

S.543
Title: An Act to reform Federal deposit insurance, protect the deposit insurance funds, recapitalize the Bank Insurance Fund, improve supervision and regulation of insured depository institutions, and for other purposes.
Sponsor: Sen Riegle, Donald W., Jr. [MI] (introduced 3/5/1991) Cosponsors (2)
Latest Major Action: 12/19/1991 Became Public Law No: 102-242.
Index terms: Banks and banking Accounting Administrative fees Cost control Credit Deposit insurance Depressed areas and other 110 terms

Adams (D-WA), **Nay** Akaka (D-HI), **Yea** Bentsen (D-TX), **Yea** Biden (D-DE), **Yea** Bond (R-MO), **Yea** Bradley (D-NJ), **Nay** Conrad (D-ND), **Nay ……**

Topics Discovered (U.S. Senate)

Mixture of Unigrams

Education	Energy	Military Misc.	Economic
education	energy	government	federal
school	power	military	labor
aid	water	foreign	insurance
children	nuclear	tax	aid
drug	gas	congress	tax
students	petrol	aid	business
elementary	research	law	employee
prevention	pollution	policy	care

Group-Topic Model

Education + Domestic	Foreign	Economic	Social Security + Medicare
education	foreign	labor	social
school	trade	insurance	security
federal	chemicals	tax	insurance
aid	tariff	congress	medical
government	congress	income	care
tax	drugs	minimum	medicare
energy	communicable	wage	disability
research	diseases	business	assistance

Groups Discovered (US Senate)

Groups from topic ***Education + Domestic***

Group 1	Group 3	Group 4
73 Republicans	Cohen(R-ME)	Armstrong(R-CO)
Krueger(D-TX)	Danforth(R-MO)	Garn(R-UT)
Group 2	Durenberger(R-MN)	Humphrey(R-NH)
90 Democrats	Hatfield(R-OR)	McCain(R-AZ)
Chafee,L.(R-RI)	Heinz(R-PA)	McClure(R-ID)
Jeffords(I-VT)	Jeffords(R-VT)	Roth(R-DE)
	Kassebaum(R-KS)	Symms(R-ID)
	Packwood(R-OR)	Wallop(R-WY)
	Specter(R-PA)	Brown(R-CO)
	Snowe(R-ME)	DeWine(R-OH)
	Collins(R-ME)	Thompson(R-TN)
		Fitzgerald(R-IL)
		Voinovich(R-OH)
		Miller(D-GA)
		Coleman(R-MN)

Senators Who Change Coalition the most Dependent on Topic

Senator	Group Switch Index
Shelby(D-AL)	0.6182
Heflin(D-AL)	0.6049
Voinovich(R-OH)	0.6012
Johnston(D-LA)	0.5878
Armstrong(R-CO)	0.5747

e.g. Senator Shelby (D-AL) votes
with the Republicans on Economic
with the Democrats on Education + Domestic
with a small group of maverick Republicans on Social Security + Medicaid

Dataset #2:

The UN General Assembly

- Voting records of the UN General Assembly (1990 - 2003)
- A country may choose to vote *Yes*, *No* or *Abstain*
- 931 resolutions with text attributes (titles)
- 192 countries in total
- Also experiments later with resolutions from 1960-2003

Vote on Permanent Sovereignty of Palestinian People, 87th plenary meeting

The draft resolution on permanent sovereignty of the Palestinian people in the occupied Palestinian territory, including Jerusalem, and of the Arab population in the occupied Syrian Golan over their natural resources (document A/54/591) was adopted by a recorded vote of 145 in favour to 3 against with 6 abstentions:

In favour: Afghanistan, Argentina, Belgium, Brazil, Canada, China, France, Germany, India, Japan, Mexico, Netherlands, New Zealand, Pakistan, Panama, Russian Federation, South Africa, Spain, Turkey, and other 126 countries.
Against: Israel, Marshall Islands, United States.
Abstain: Australia, Cameroon, Georgia, Kazakhstan, Uzbekistan, Zambia.

Topics Discovered (UN)

Mixture of Unigrams

Everything Nuclear	Human Rights	Security in Middle East
nuclear	rights	occupied
weapons	human	israel
use	palestine	syria
implementation	situation	security
countries	israel	calls

Group-Topic Model

Nuclear Non-proliferation	Nuclear Arms Race	Human Rights
nuclear	nuclear	rights
states	arms	human
united	prevention	palestine
weapons	race	occupied
nations	space	israel

Groups Discovered (UN)

The countries list for each group are ordered by their 2005 GDP (PPP) and only 5 countries are shown in groups that have more than 5 members.

GROUP ↓	Nuclear Arsenal	Human Rights	Nuclear Arms Race
	nuclear states united weapons nations	rights human palestine occupied israel	nuclear arms prevention race space
1	Brazil Columbia Chile Peru Venezuela	Brazil Mexico Columbia Chile Peru	UK France Spain Monaco East-Timor
2	USA Japan Germany UK... Russia	Nicaragua Papua Rwanda Swaziland Fiji	India Russia Micronesia
3	China India Mexico Iran Pakistan	USA Japan Germany UK... Russia	Japan Germany Italy... Poland Hungary
4	Kazakhstan Belarus Yugoslavia Azerbaijan Cyprus	China India Indonesia Thailand Philippines	China Brazil Mexico Indonesia Iran
5	Thailand Philippines Malaysia Nigeria Tunisia	Belarus Turkmenistan Azerbaijan Uruguay Kyrgyzstan	USA Israel Palau

Groups and Topics, Trends over Time (UN)

Time Period	Topic 1	Topic 2	Topic 3	Group distributions for Topic 3: Group 1	Group2	Group3	Group4	Group5
60-75	**Nuclear** operative general nuclear power	**Procedure** committee amendment assembly deciding	**Africa Indep.** calling right africa self	India Indonesia Iran Thailand Philippines	USA Japan UK France Italy	Argentina Colombia Chile Venezuela Dominican	USSR Poland Hungary Bulgaria Belarus	Turkey
65-80	**Independence** territories independence self colonial	**Finance** budget appropriation contribution income	**Weapons** nuclear UN international weapons	Cuba Albania	India Indonesia Pakistan Saudi Egypt	Algeria Iraq Syria Libya Afganistan	USSR Poland Hungary Bulgaria Belarus	USA Japan UK France Italy
70-85	**N. Weapons** nuclear international UN human	**Israel** israel measures hebron expelling	**Rights** africa territories south right	Mexico Indonesia Iran Thailand Philippines	China	USA Japan UK France Italy	Brazil Turkey Argentina Colombia Chile	India USSR Poland Vietnam Hungary
75-90	**Rights** south africa israel rights	**Israel/Pal.** israel arab occupied palestine	**Disarmament** UN international nuclear disarmament	Mexico Indonesia Iran Thailand Philippines	USA Japan UK France USSR	Algeria Vietnam Iraq Syria Libya	China Brazil Argentina Colombia Chile	India
80-95	**Disarmament** nuclear US disarmament international	**Conflict** need israel palestine secretary	**Pal. Rights** rights palestine israel occupied	USA Israel	China India Russia Spain Hungary	Japan UK France Italy Canada	Guatemala St Vincent Dominican	Malawi
85-00	**Weapons** nuclear weapons use international	**Rights** rights human fundamental freedoms	**Israel/Pal.** israeli palestine occupied disarmament	Poland Czech R. Hungary Bulgaria Albania	China India Brazil Mexico Indonesia	USA Japan UK France Italy	Russia Argentina Ukraine Belarus Malta	Cameroon Congo Ivory C. Liberia

Social Networks in Research Literature

- Better understand structure of our own research area.
- Structure helps us learn a new field.
- Aid collaboration
- Map how ideas travel through social networks of researchers.

- Aids for hiring and finding reviewers!

Traditional Bibliometrics

- Analyses a small amount of data
  (e.g. 19 articles from a single issue of a journal)

- Uses “journal” as a proxy for “research topic”
  (but there is no journal for information extraction)

- Uses impact measures almost exclusively
  based on simple citation counts.

**How can we use topic models
to create new, interesting impact measures?**

Our Data

- Over 1 million research papers, gathered as part of *Rexa.info* portal.
- Cross linked references / citations.

Finding Topics with TNG

Traditional unigram LDA run on 1 million titles / abstracts (200 topics)

...select ~300k papers on ML, NLP, robotics, vision...

Find 200 TNG topics among those papers.

Topic	Topic Unigrams
Web1 (98)	web information search digital user library users pages content libraries
Web2 (156)	web semantic ontology services world wide based ontologies hypermedia metadata
Computer Vision (5)	recognition object face tracking objects based system image video human
Game Theory (111)	decision making utility equilibrium games theory game choice preferences model
System (160)	system performance communication operating parallel implementation network applications message high

Topic	Topic Unigrams and Ngrams
Digital Libraries (102)	digital electronic library metadata access
	"digital libraries" "digital library" "electronic commerce" "dublin core" "cultural heritage"
Web Pages (129)	web site pages page www sites
	"world wide web" "web pages" "web sites" "web site" "world wide"
Ontologies (186)	semantic ontology ontologies rdf semantics meta
	"semantic web" "description logics" "rdf schema" "description logic" "resource description framework"
Web Services (184)	web services service xml business
	"web services" "web service" "markup language" "xml documents" "xml schema"

Topical Bibliometric Impact Measures

[Mann, Mimno, McCallum, 2006]

- Topical Citation Counts
- Topical Impact Factors
- Topical Longevity
- Topical Diversity
- Topical Precedence
- Topical Transfer

Topical Diversity

Entropy of the topic distribution among papers that cite this paper (this topic).

Topic	Impact Diversity
Computer Vision (49)	2.95
Speech Recognition (120)	3.09
Graphics (9)	3.21
Mobile Robots (22)	3.31
Machine Translation (96)	3.32
Digital Libraries (102)	3.77
Probability (89)	4.5
Finite Automata (66)	4.55
Probabilistic Modeling (3)	4.55
Pattern Recognition (125)	4.57
Simulated Annealing (52)	4.59

Low Diversity

High Diversity

Topical Diversity

Can also be measured on particular papers...

Topical Diversity	Citations	Title
4.00	618	A tutorial on hidden Markov models and selected applications in speech processing
3.80	138	The self-organizing map
3.77	163	Hierarchical mixtures of experts and the EM algorithm
3.74	65	Quantifying Inductive Bias: AI Learning Algorithms and ...
3.74	144	Knowledge Acquisition via Incremental Conceptual Clustering
3.73	155	A Tutorial on Learning With Bayesian Networks
3.72	244	Term-Weighting Approaches in Automatic Text Retrieval
3.71	294	Finding Structure in Time
3.7	173	An introduction to hidden Markov models
3.7	132	Nearest neighbor pattern classification

Topical Precedence

"Early-ness"

Within a topic, what are the earliest papers that received more than *n* citations?

<u>Information Retrieval:</u>

On Relevance, Probabilistic Indexing and Information Retrieval,
Kuhns and Maron (1960)

Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems,
Cooper (1968)

Relevance feedback in information retrieval,
Rocchio (1971)

Relevance feedback and the optimization of retrieval effectiveness,
Salton (1971)

New experiments in relevance feedback,
Ide (1971)

Automatic Indexing of a Sound Database Using Self-organizing Neural Nets,
Feiten and Gunzel (1982)

Topical Precedence

“Early-ness”

Within a topic, what are the earliest papers that received more than *n* citations?

<u>Speech Recognition:</u>

Some experiments on the recognition of speech, with one and two ears,
E. Colin Cherry (1953)

Spectrographic study of vowel reduction,
B. Lindblom (1963)

Automatic Lipreading to enhance speech recognition,
Eric D. Petajan (1965)

Effectiveness of linear prediction characteristics of the speech wave for...,
B. Atal (1974)

Automatic Recognition of Speakers from Their Voices,
B. Atal (1976)

Topical Transfer

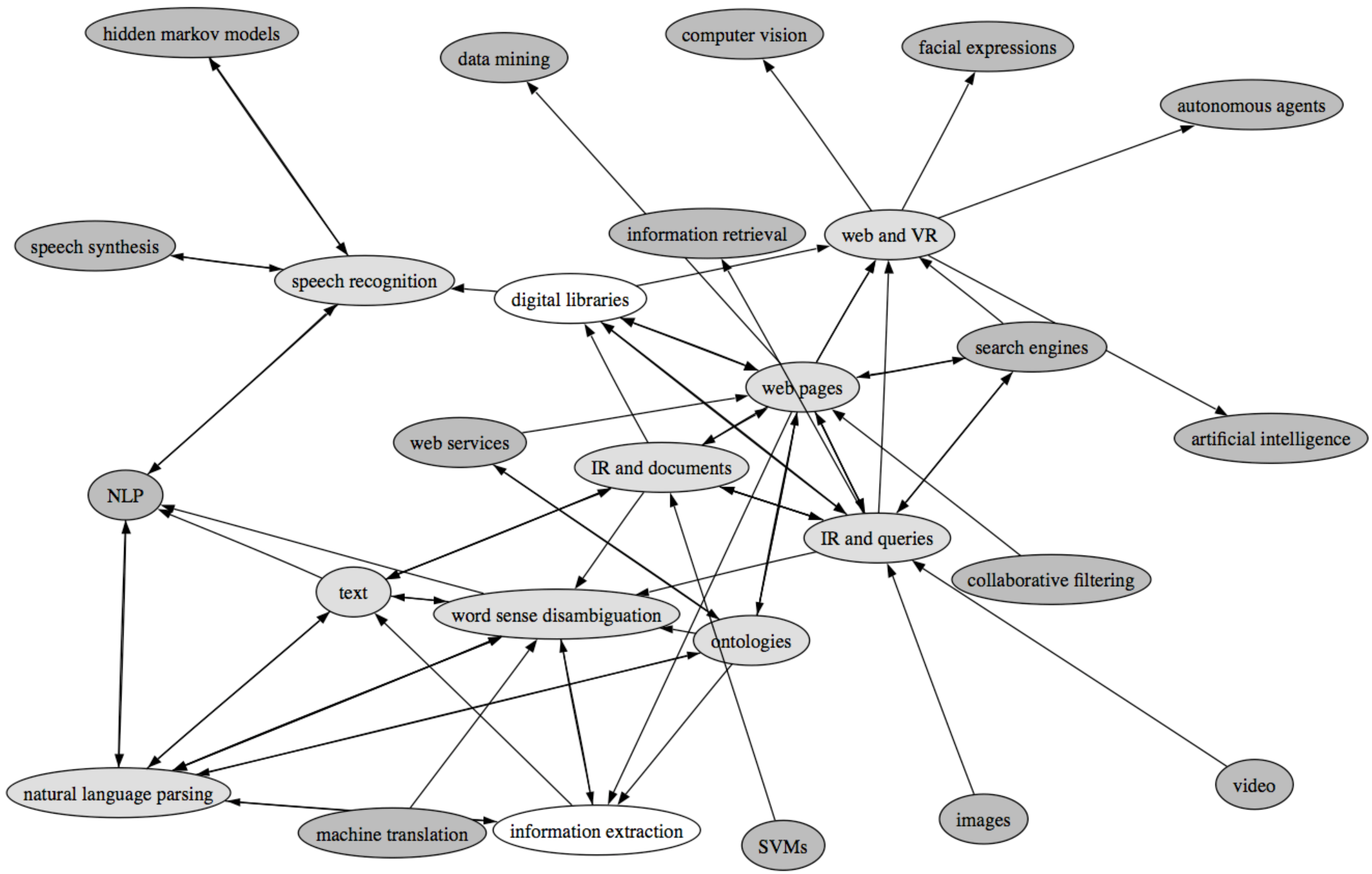
Transfer from Digital Libraries to other topics

Other topic	**Cit's**	**Paper Title**
Web Pages	31	*Trawling the Web for Emerging Cyber-Communities,* Kumar, Raghavan,... 1999.
Computer Vision	14	*On being 'Undigital' with digital cameras: extending the dynamic...*
Video	12	*Lessons learned from the creation and deployment of a terabyte digital video*
Graphs	12	*Trawling the Web for Emerging Cyber-Communities*
Web Pages	11	*WebBase: a repository of Web pages*

Topical Transfer

Citation counts from one topic to another.

Map "producers and consumers"



Outline

Social Network Analysis with Topic Models

- ✔ Role Discovery (Author-Recipient-Topic Model, ART)
- ✔ Group Discovery (Group-Topic Model, GT)
- ✔ Enhanced Topic Models
  - ✔ – Correlations among Topics (Pachinko Allocation, PAM)
  - ✔ – Time Localized Topics (Topics-over-Time Model, TOT)
  - ✔ – Markov Dependencies in Topics (Topical N-Grams Model, TNG)
- ✔ Bibliometric Impact Measures enabled by Topics

Multi-Conditional Mixtures

Topic Model Musings

- 3 years ago Latent Dirichlet Allocation appeared as a complex innovation ...but now these methods & mechanics are well-understood.

- Innovation now is to understand data and modeling needs, how to structure a new model to capture these.