Probabilistic Approaches for Modeling Text Structure and their Application to Text-to-Text Generation

Regina Barzilay
(MIT)
joint work with Branavan, Harr Chen,
Mirella Lapata, Lillian Lee, Christina Sauper

Guess the Author…

Active networks and virtual machines have a long history of collaborating in this manner. The basic tenet of this solution is the refinement of Scheme. The disadvantage of this type of approach, however, is that public-private key pair and red-black trees are rarely incompatible.

SCIgen: An Automatic CS Paper Generator

- An output of a system that automatically generates scientific papers (Stribling et al., 2005):

> Active networks and virtual machines have a long history of collaborating in this manner. The basic tenet of this solution is the refinement of Scheme. The disadvantage of this type of approach, however, is that public-private key pair and red-black trees are rarely incompatible.

- The paper was accepted to a conference (not ACL!)

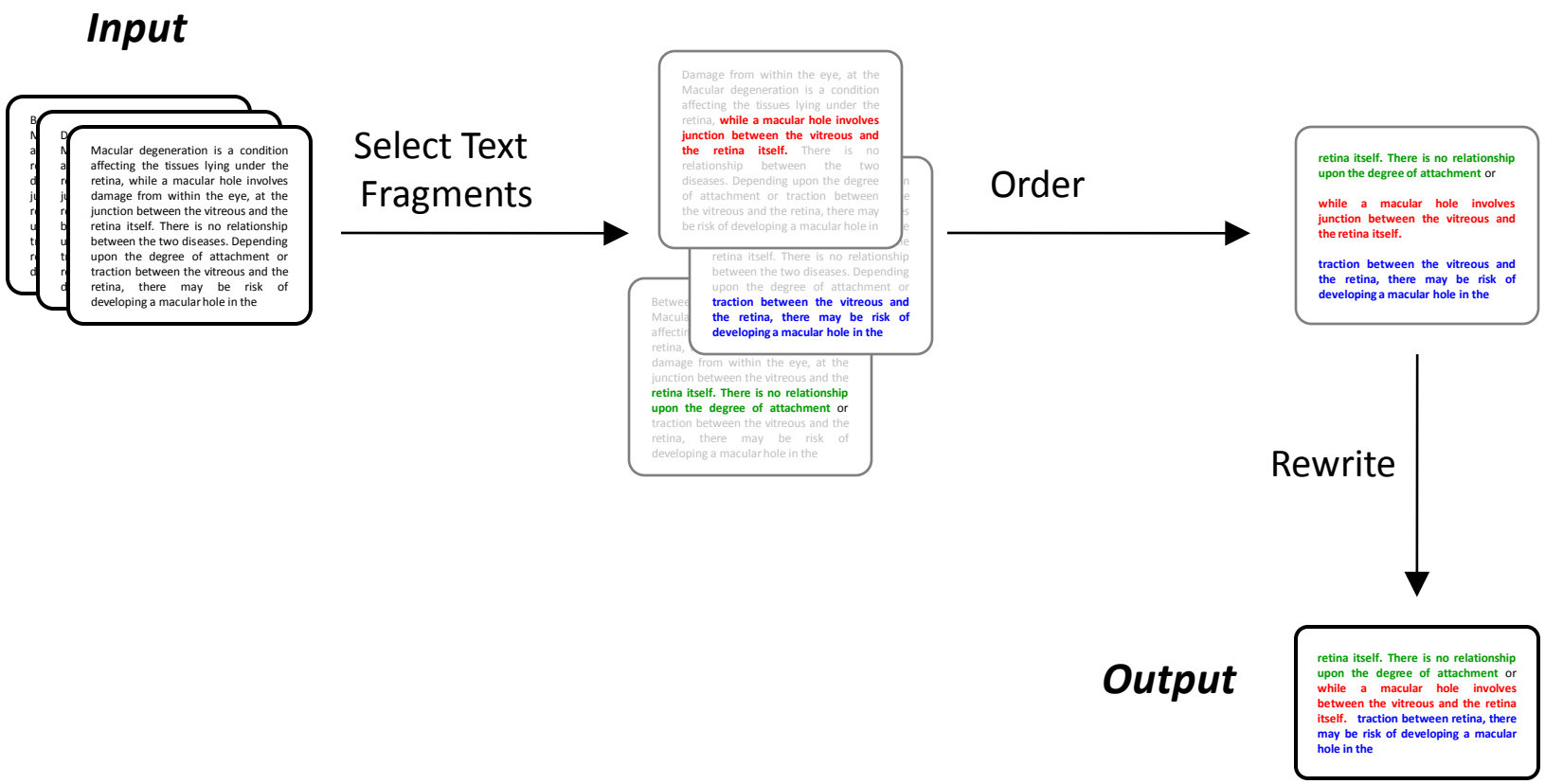
Less Exotic Example

- An output of a state-of-the-art multidocument summarization system

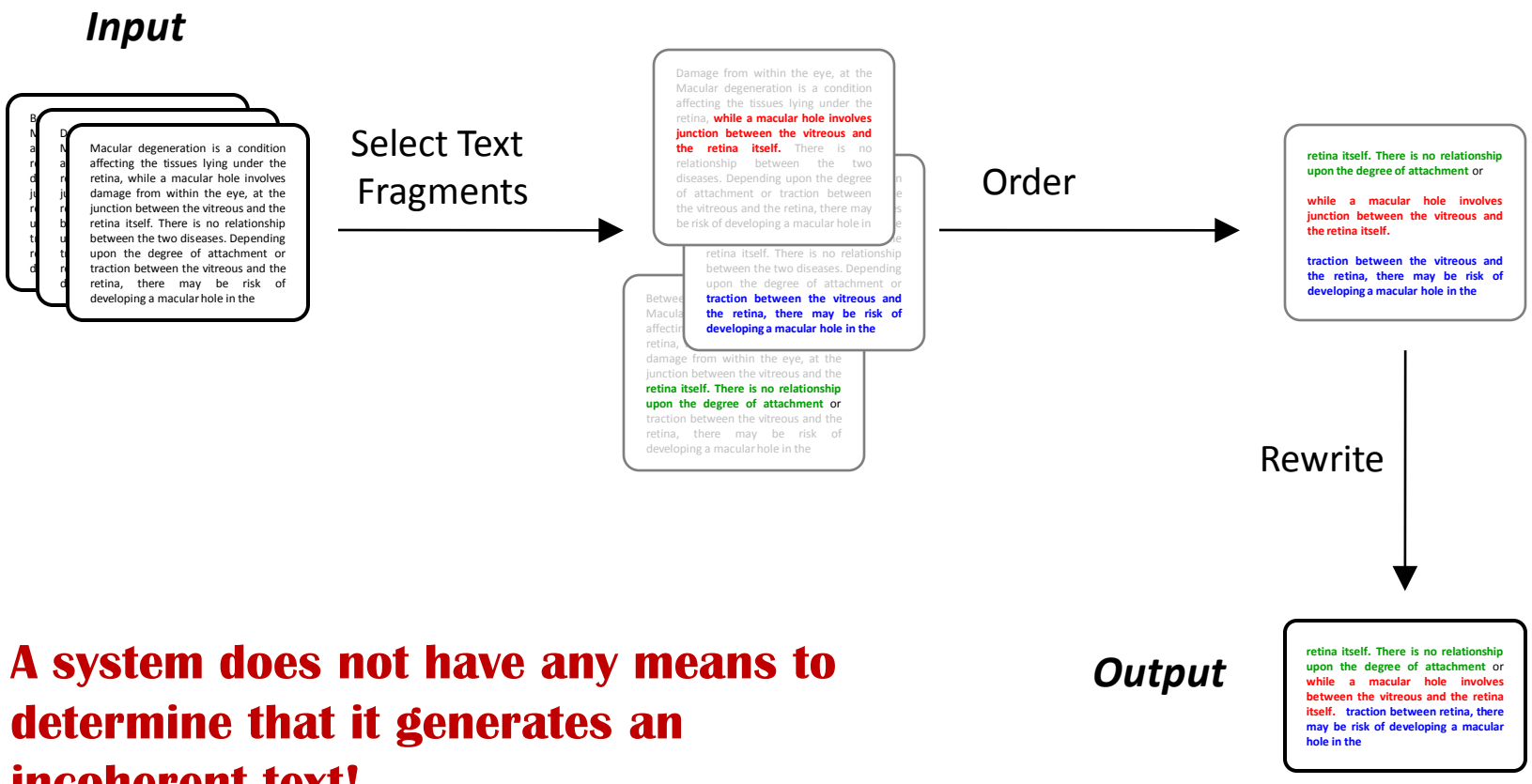
> Newspapers reported Wednesday that three top Libyan officials have been tried and jailed in the Lockerbie case. The sanctions were imposed to force Libyan leader Moammar Gadhafi to turn the men over. *Louis Farrakhan congratulated Gadhafi on his recovery from a hip injury.*

- A system does not have any means to determine that it generates an incoherent text

Text-to-text Generation



Text-to-text Generation



A system does not have any means to determine that it generates an incoherent text!

Linguistic Foundations: Discourse Theory

- Centering Theory (Grosz, Joshi & Weinstein, 1983)
- Rhetorical Structure Theory (Mann & Thompson, 1988)
- Content Schema (Bartlett, 1932)
- …

Current Solutions

- Work hard: manually encode all the constraints
  - Expensive
  - Not scalable
- Ignore coherence, and hope for the best
  - Incoherent output

Our goal: linguistically-motivated text models without manual crafting

The Goal

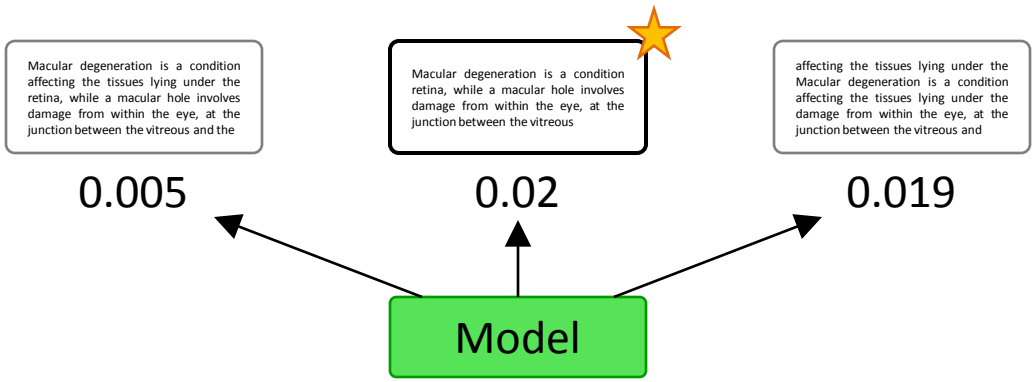
Our goal: linguistically-motivated text models without manual crafting

"various types of [word] recurrence patterns seem to characterize various types of discourse'' (Harris, 1982)

Outline

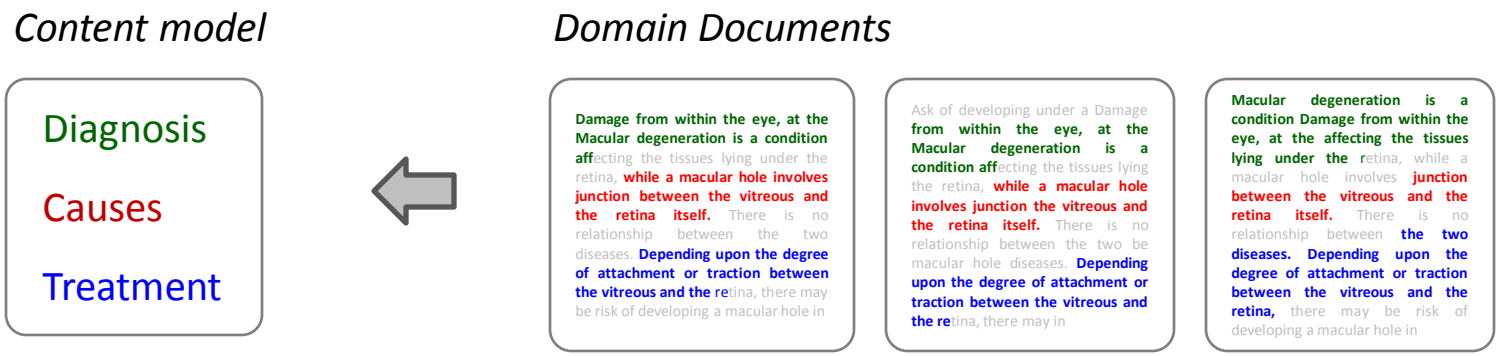
- Application of text models
- Content Models (NAACL 2004, NAACL 2009)
- Coherence Models (ACL 2005, CL 2008)

Generating Coherent Text (1)



- Model assesses the likelihood that the input is well-formed
- It ranks candidate outputs based on the model scores
- Analogous to the use of language model in ASR and MT

Generating Coherent Text (2)



- Specifies patterns in topic occurrence and their ordering
- Each topic is represented as a distribution over words

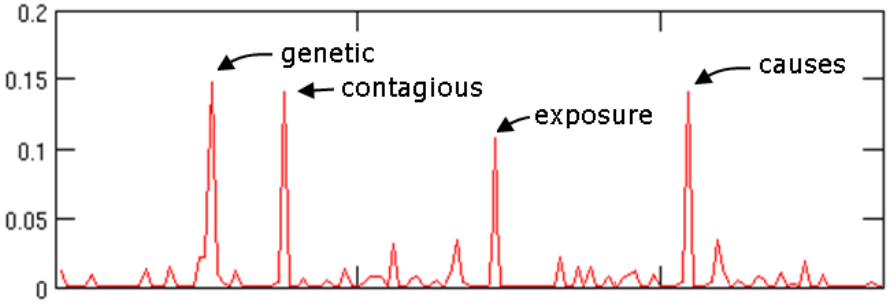
Generating Wikipedia Articles

- Induce a content model from existing Wikipedia articles in the target domain:
  - Query for Internet search
    (e.g., *"Macular Hole" treatment*)
  - Extractor of relevant content
    (trained on existing articles)

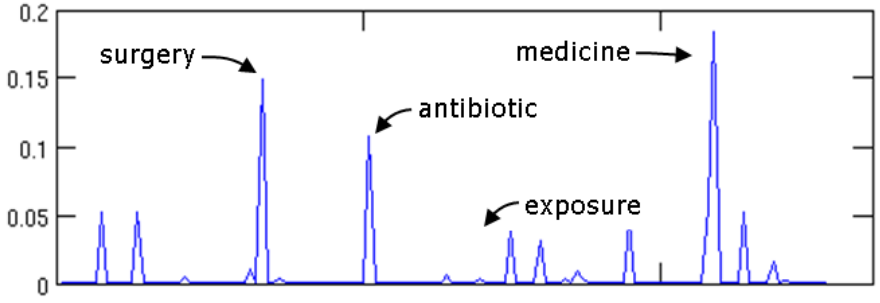
Diagnosis

Causes

Treatment



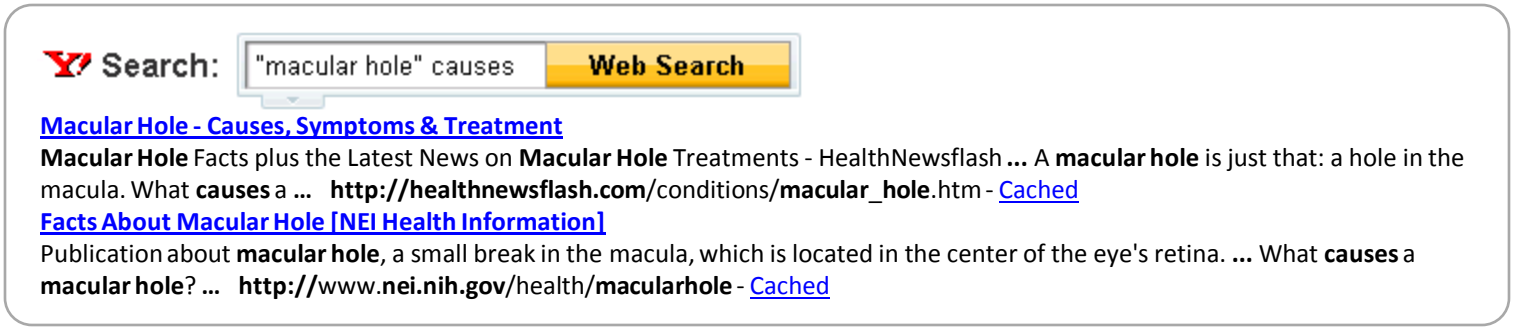
Causes



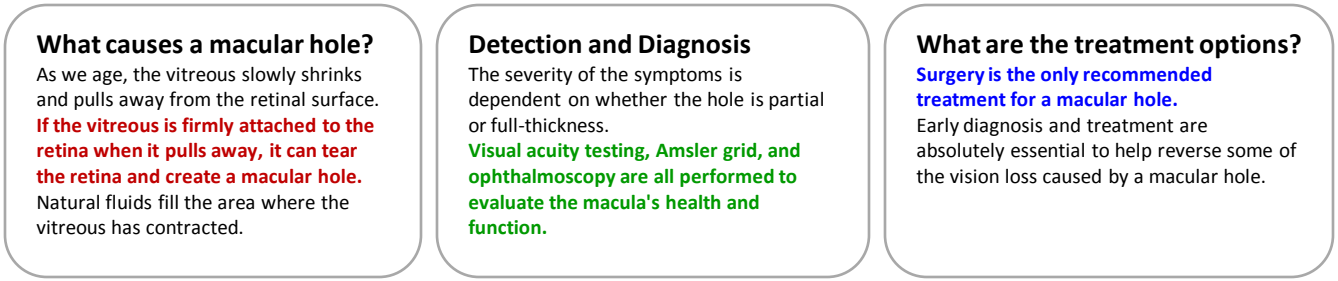
Treatment

Generating Wikipedia Articles (2)

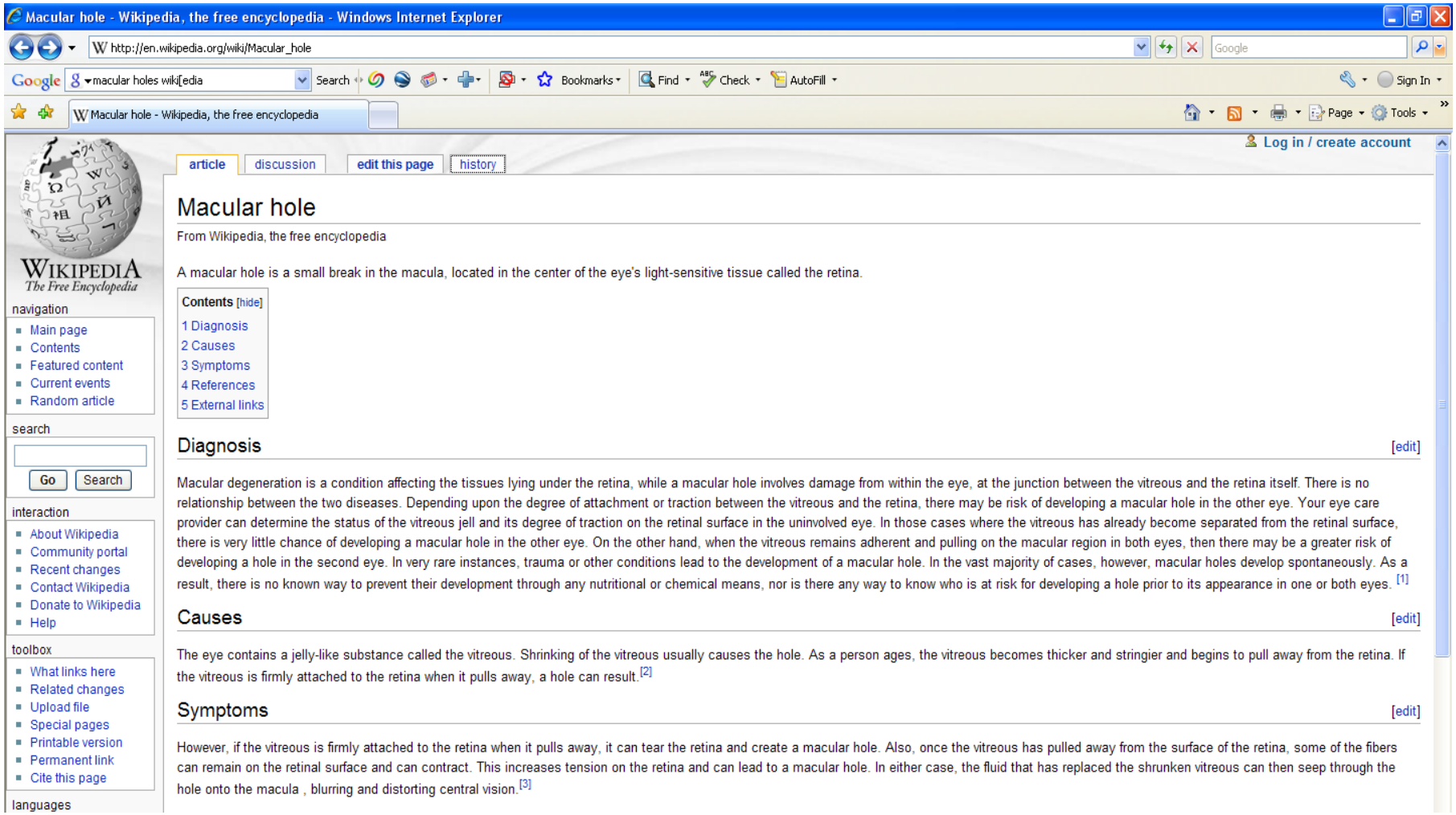
- Retrieve web pages using specified queries



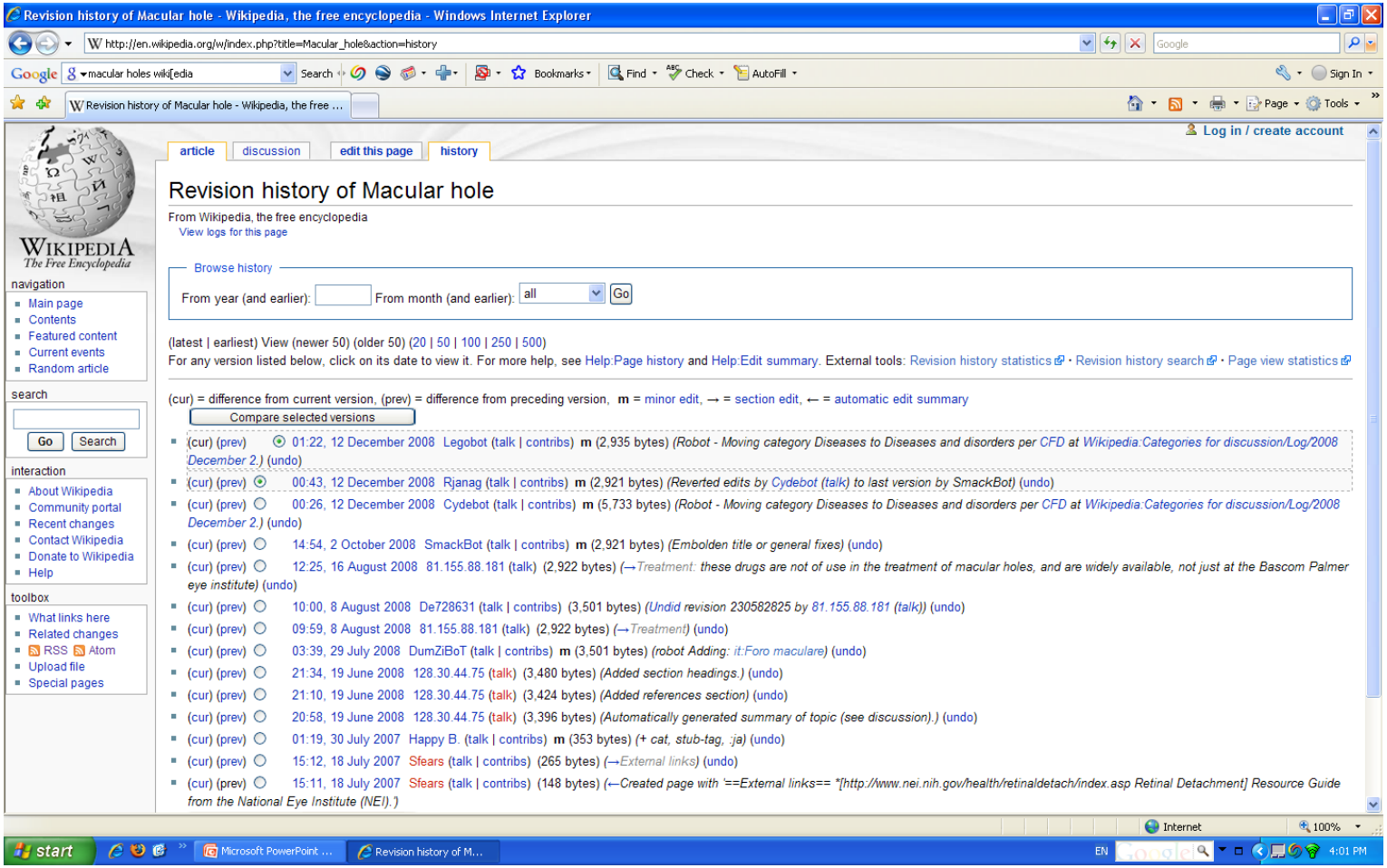
- Jointly select and order excerpts



Example Output from Wikipedia



Readers' Reaction

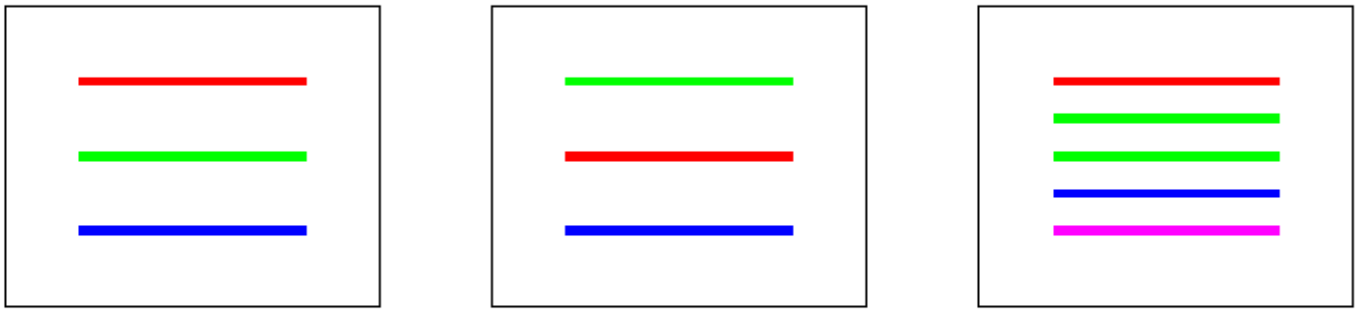


More Automatically-Generated Wikipedia Articles

http://en.wikipedia.org/wiki/Felty_syndrome
http://en.wikipedia.org/wiki/Cogan_syndrome
http://en.wikipedia.org/wiki/Blue_rubber_bleb_nevus_syndrome
http://en.wikipedia.org/wiki/Barraquer-Simons_syndrome
http://en.wikipedia.org/wiki/Chediak-Higashi_syndrome
http://en.wikipedia.org/wiki/Ancylostomiasis
http://en.wikipedia.org/wiki/Amyoplasia
http://en.wikipedia.org/wiki/Hemorrhagic_cystitis
http://en.wikipedia.org/wiki/Hartnup_disease
http://en.wikipedia.org/wiki/Heterotopic_ossification
http://en.wikipedia.org/wiki/Hypophosphatasia
http://en.wikipedia.org/wiki/Vestibular_neuronitis
http://en.wikipedia.org/wiki/3-M_syndrome
http://en.wikipedia.org/wiki/Macular_hole
http://en.wikipedia.org/wiki/Hermansky-Pudlak_syndrome

Ordering Task 1

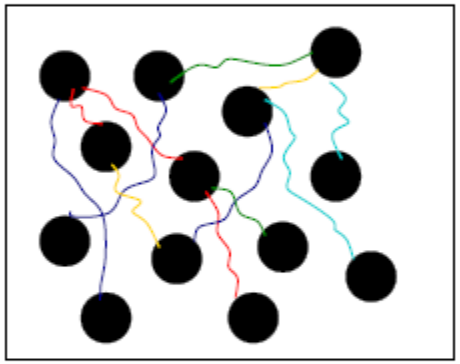
Input: N alternative text realizations



Task: Find the most coherent alternative

Ordering Task 2

Input: N information-bearing items



Task: Organize input items into a text based on their semantic relations

Information Ordering

(a) During a third practice forced landing, with the landing gear extended, the CFI took over the controls.

(b) The certified flight instructor (CFI) and the private pilot, her husband, had flown a previous flight that day and practiced maneuvers at altitude.

(c) The private pilot performed two practice power off landings from the downwind to runway 18.

(d) When the airplane developed a high sink rate during the turn to final, the CFI realized that the airplane was low and slow.

(e) After a refueling stop, they departed for another training flight.

Information Ordering

(b) The certified flight instructor (CFI) and the private pilot, her husband, had flown a previous flight that day and practiced maneuvers at altitude.

(e) After a refueling stop, they departed for another training flight.

(c) The private pilot performed two practice power off landings from the downwind to runway 18.

(a) During a third practice forced landing, with the landing gear extended, the CFI took over the controls.

(d) When the airplane developed a high sink rate during the turn to final, the CFI realized that the airplane was low and slow.

Outline

- Applications of Text Models
- Coherence Models (ACL 2005, CL 2008)
- Content Models (NAACL 2004, NAACL 2009)

Content Models

Content models represent topics and their ordering in a domain text

> Domain: newspaper articles on earthquakes
>
> Topics: “strength,” “location,” “casualties,” . . .
>
> Order: “casualties” prior to “rescue efforts”

Learning Content Structure

- Our goal: learn content structure from un-annotated texts via analysis of word distribution patterns
- The success of the distributional approach depends on the existence of recurrent patterns.
  - Linguistics: domain-specific texts tend to exhibit high similarity (Wray, 2002)
  - Cognitive psychology: formulaic text structure facilitates readers' comprehension(Bartlett, 1932)

Patterns in Content Organization

TOKYO (AP) A moderately strong earthquake rattled northern Japan early Wednesday, the Central Meteorological Agency said. There were no immediate reports of casualties or damage. The quake struck at 6:06 am (2106 GMT) 60 kilometers (36 miles) beneath the Pacific Ocean near the northern tip of the main island of Honshu. . . .

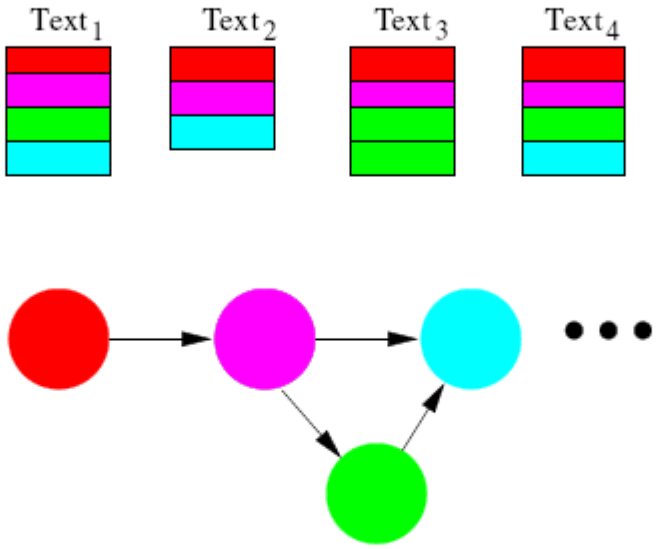
ATHENS, Greece (AP) A strong earthquake shook the Aegean Sea island of Crete on Sunday but caused no injuries or damage. The quake had a preliminary magnitude of 5.2 and occurred at 5:28 am (0328 GMT) on the sea floor 70 kilometers (44 miles) south of the Cretan port of Chania. . . .

Patterns in Content Organization

TOKYO (AP) A moderately strong earthquake rattled northern Japan early Wednesday, the Central Meteorological Agency said. There were no immediate reports of casualties or damage. The quake struck at 6:06 am (2106 GMT) 60 kilometers (36 miles) beneath the Pacific Ocean near the northern tip of the main island of Honshu. . . .

ATHENS, Greece (AP) A strong earthquake shook the Aegean Sea island of Crete on Sunday but caused no injuries or damage. The quake had a preliminary magnitude of 5.2 and occurred at 5:28 am (0328 GMT) on the sea floor 70 kilometers (44 miles) south of the Cretan port of Chania. . . .

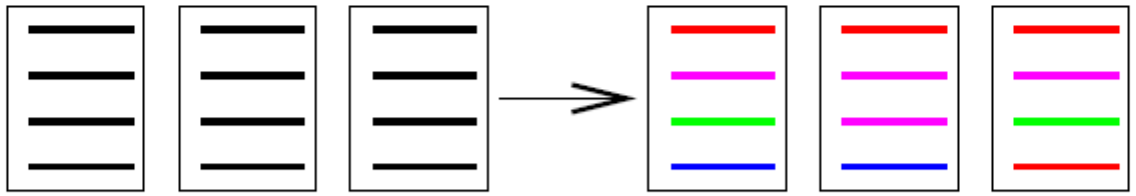
Computing Content Model



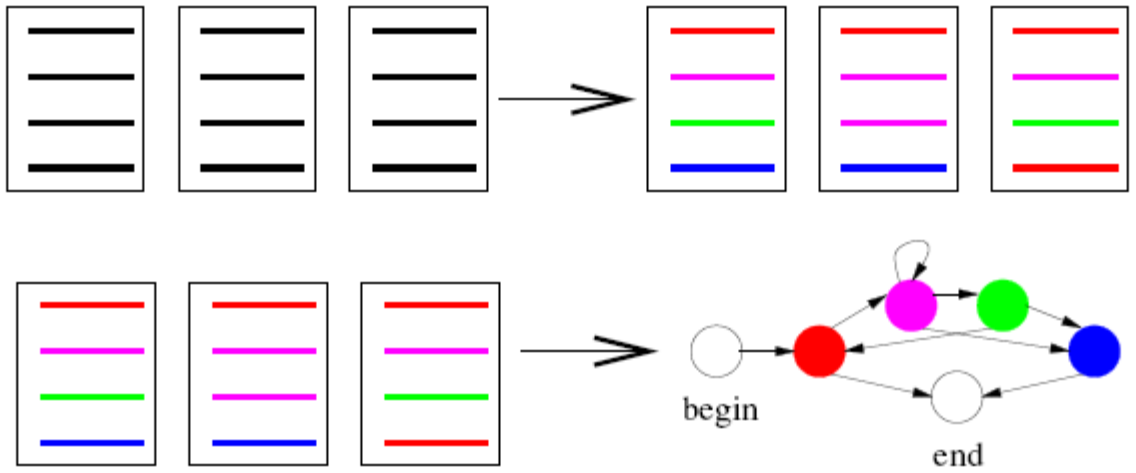
Implementation: Hidden Markov Model

- States represent topics
- State-transitions represent ordering constraints

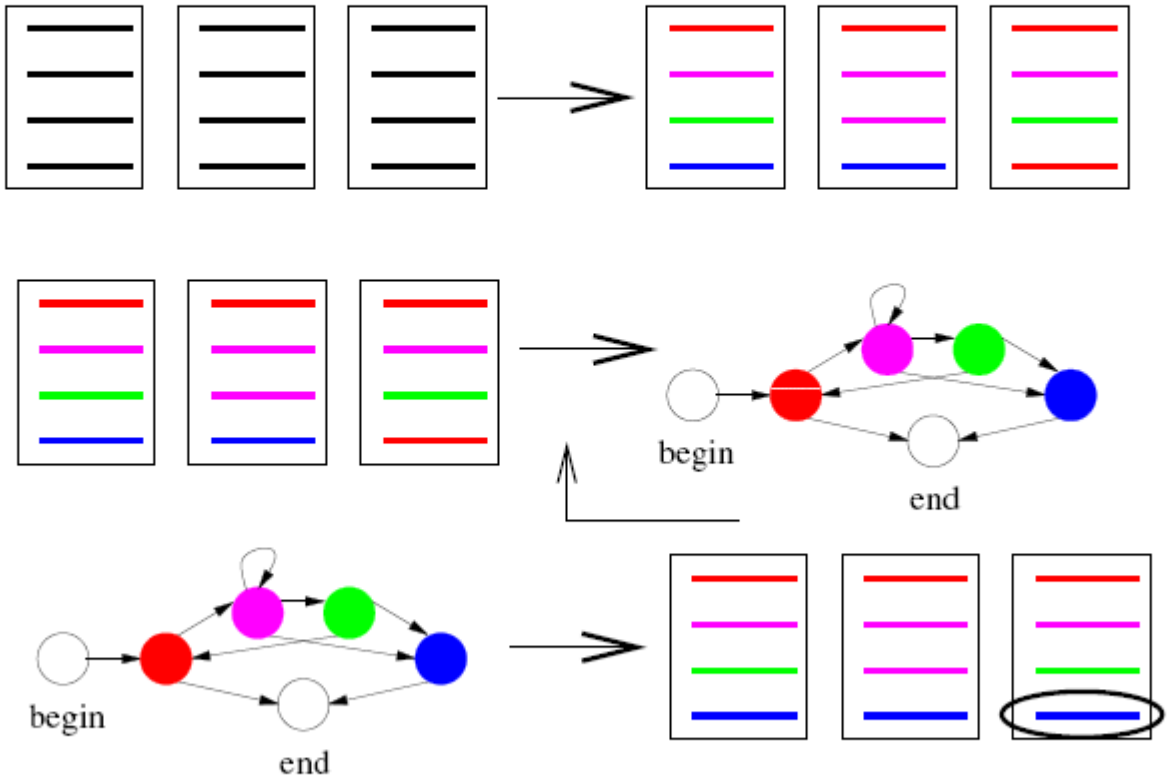
Model Induction



Model Induction



Model Induction



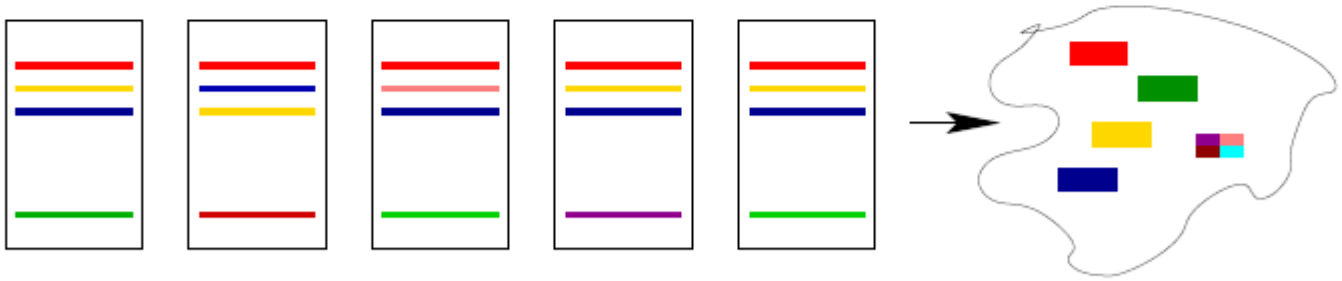
Initial Topic Induction

Agglomerative clustering with cosine similarity measure

The Athens seismological institute said the temblor's epicenter was located 380 kilometers (238 miles) south of the capital.
Seismologists in Pakistan's Northwest Frontier Province said the temblor's epicenter was about 250 kilometers (155 miles) north of the provincial capital Peshawar.
The temblor was centered 60 kilometers (35 miles) northwest of the provincial capital of Kunming, about 2,200 kilometers (1,300 miles) southwest of Beijing, a bureau seismologist said.

From Clusters to States

- Each large cluster constitutes a state
- Agglomerate small clusters into an "*insert*" state



Estimating Emission Probabilities

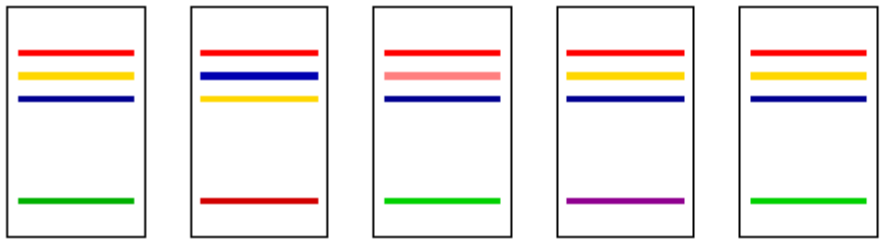
- Estimation for a "normal" state:

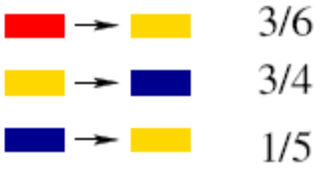
$$p_{s_i}(w'|w) \stackrel{def}{=} \frac{f_{c_i}(ww') + \delta_1}{f_{c_i}(w) + \delta_1|V|},$$

- Estimation for the "insertion" state:

$$p_{s_m}(w'|w) \stackrel{def}{=} \frac{1 - \max_{i<m} p_{s_i}(w'|w)}{\sum_{u \in V}(1 - \max_{i<m} p_{s_i}(u|w))}.$$

Estimating Transition Probabilities





$$p(s_j|s_i) = \frac{g(c_i, c_j) + \delta_2}{g(c_i) + \delta_2 m}$$

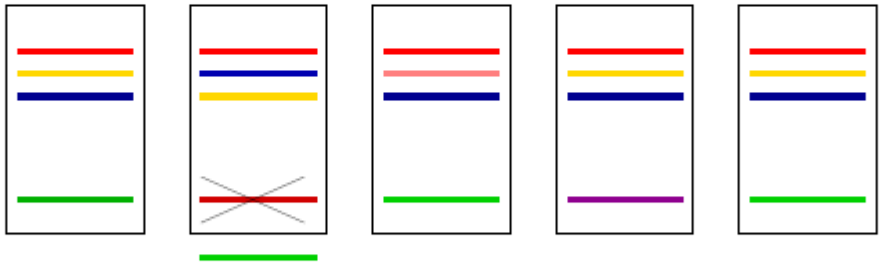
$g(c_i, c_j)$ is a number of adjacent sentences (c_i, c_j)

$g(c_i)$ is a number of sentences in c_i

Viterbi re-estimation

Goal: incorporate ordering information

- Decode the training data with Viterbi decoding



- Use the new clustering as the input to the parameter estimation procedure

Evaluation: Information Ordering

- Goal: recover the most coherent sentence ordering
- Basic set-up:
  - Input: a pair of a source document and a permutation of its sentences

  - Task: find a source document via coherence ranking
- Data: Training 4000 pairs, Testing 4000 pairs (Natural disasters and Transportation Safety Reports)

Evaluation Results: Ranking

Average number of sentences per document: 10

Model	Disasters	NTSB Reports
Content Models (HMM)	88.0	75.8
Random	50	50

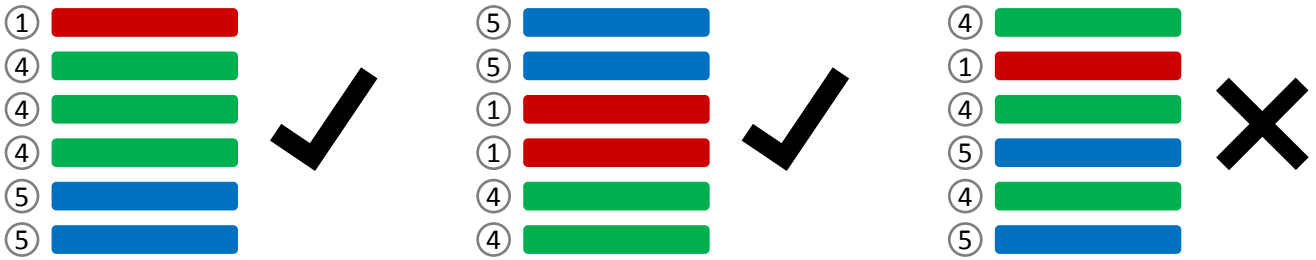
Evaluation Results: Ordering

Average number of sentences per document: 10

Model	Disasters	NTSB Reports
Content Models (HMM)	81.0	44.0
Random	<1.0	<1.0

The Need for Global Structure

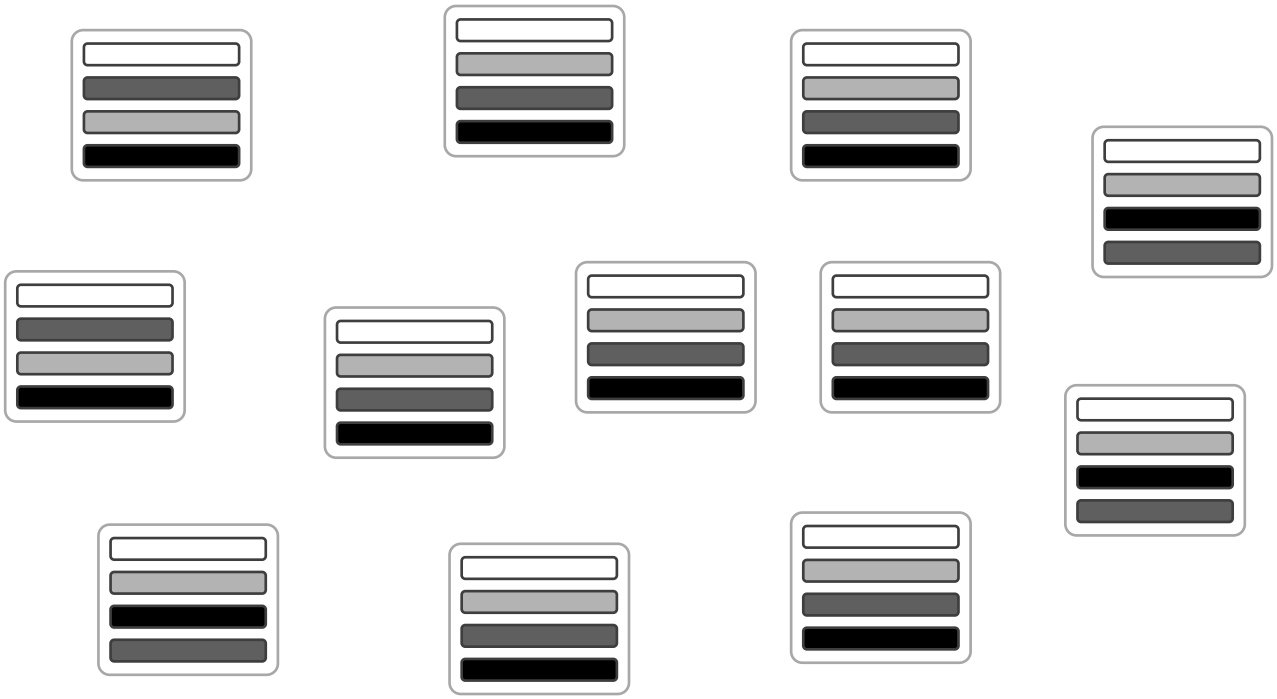
- ***Challenge:*** preserving topic contiguity
  - Topics do not repeat in disconnected sections of a document



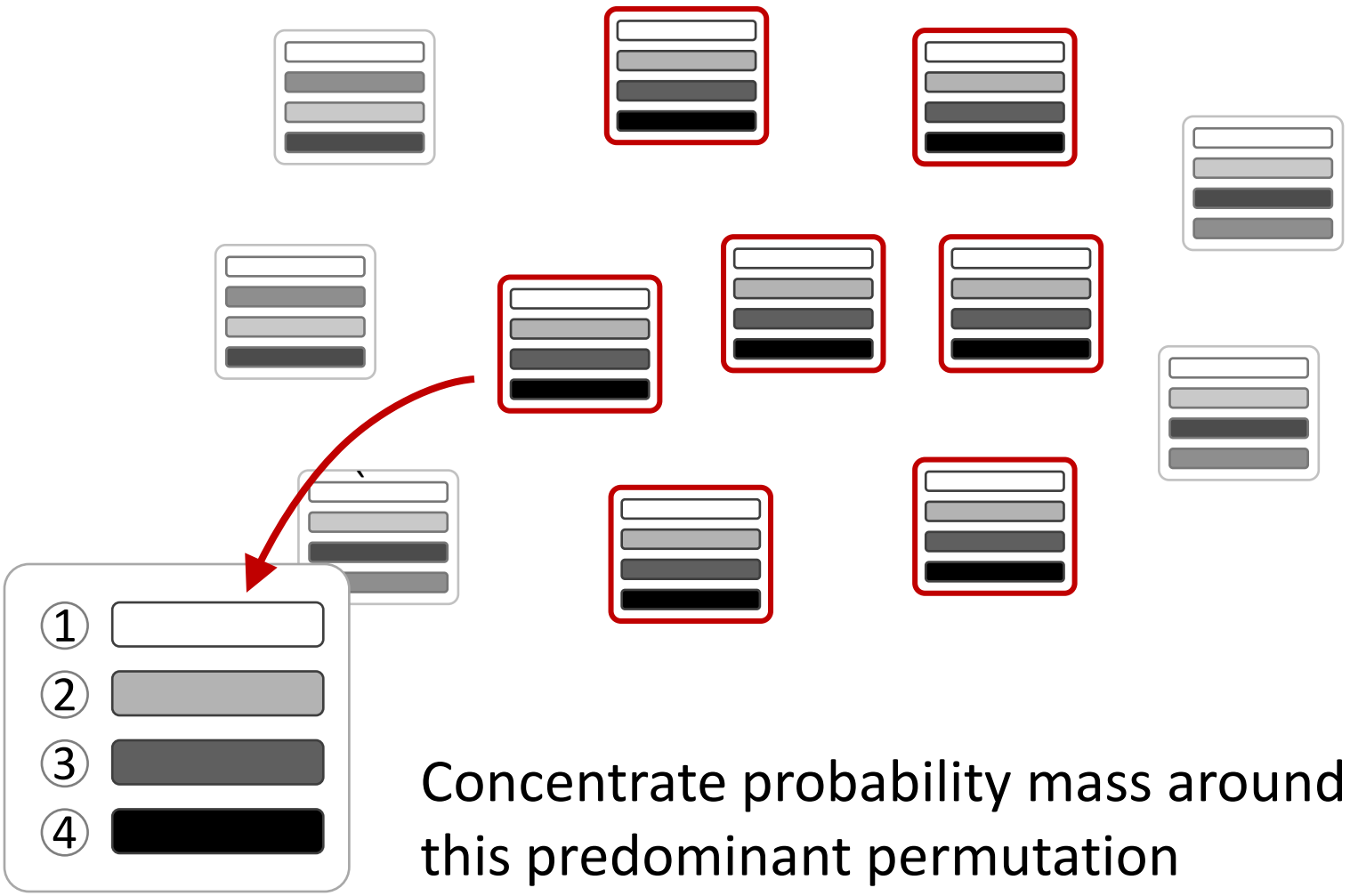
- ***Solution***: model topic structure as *permutation*

The Need for Global Structure

- ***Challenge:*** model shared structure across documents



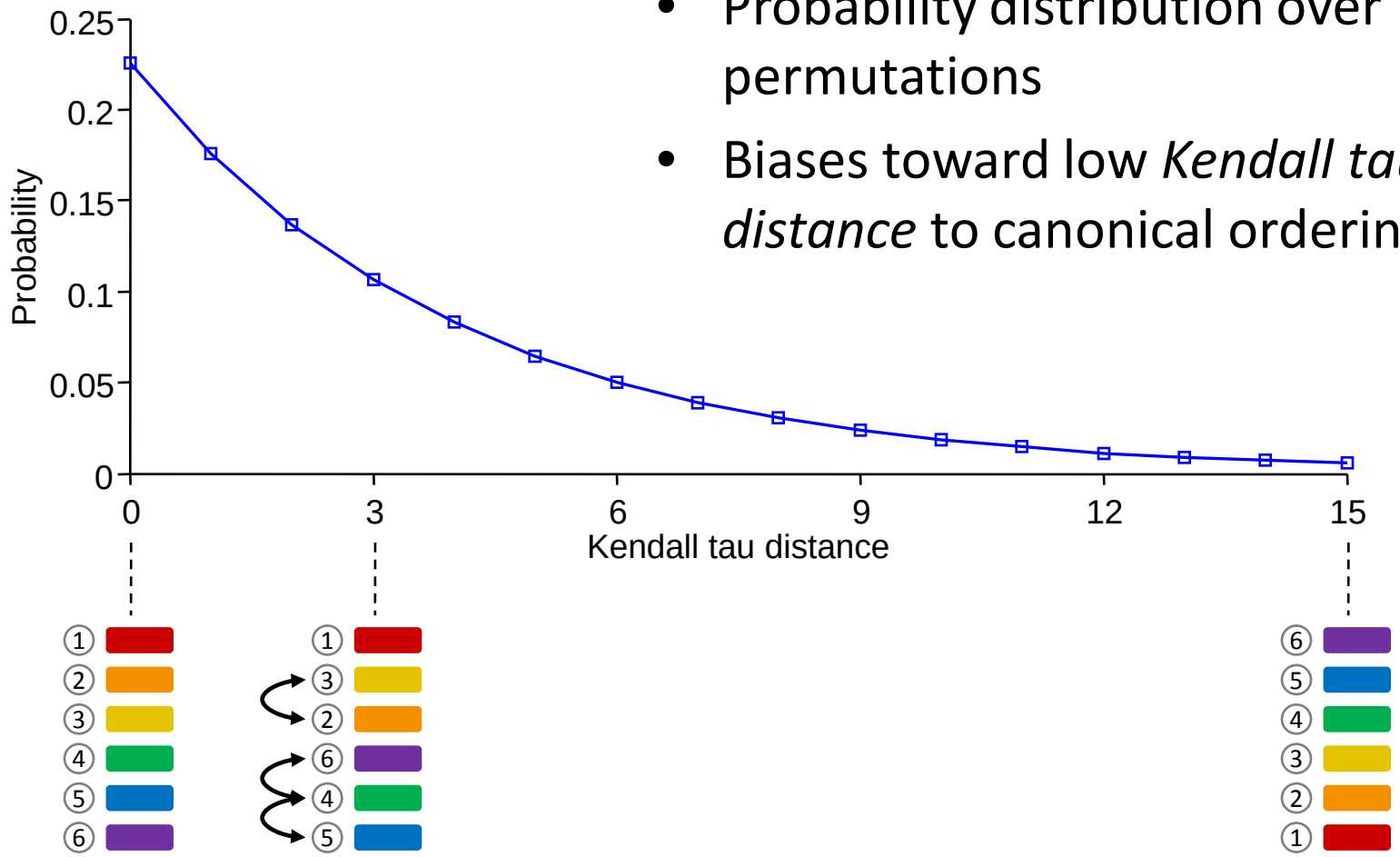
Modeling Predominant Permutations



Concentrate probability mass around this predominant permutation

The Generalized Mallows Model over Permutations

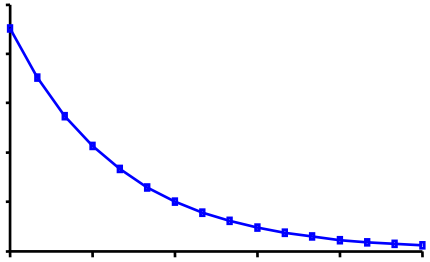
- Probability distribution over permutations
- Biases toward low *Kendall tau distance* to canonical ordering



Learning Framework

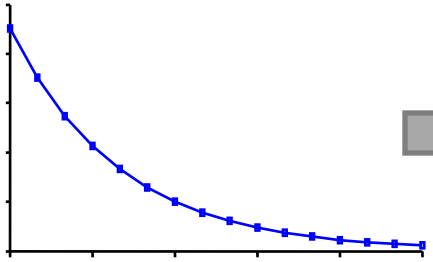
- Training
  - *Given*: Unannotated collection of documents
  - *Goal*: Estimate model parameters
    - *Predominant ordering, topic language models, etc.*
- Testing
  - *Given*: Bag of text fragments
  - *Goal*: Find maximum likelihood ordering

Generative Process

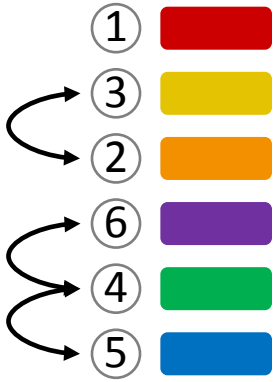


Mallows Distribution

Generative Process

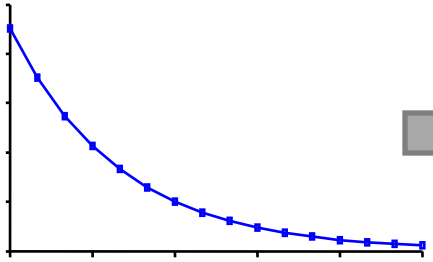


Mallows Distribution

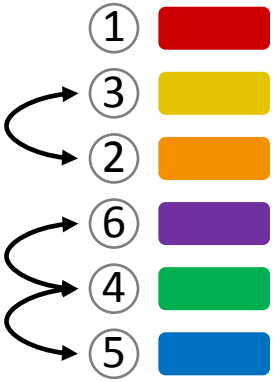


Topic Permutation

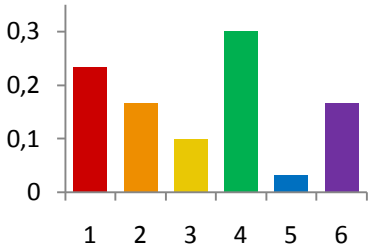
Generative Process



Mallows Distribution

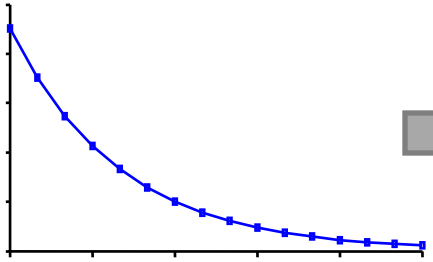


Topic Permutation

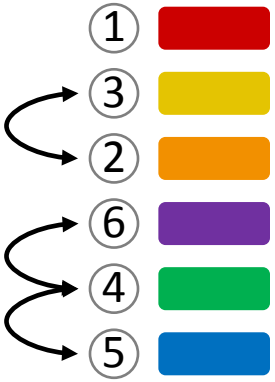


Topic Count Distribution

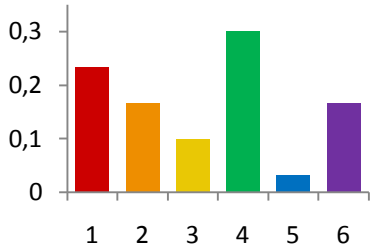
Generative Process



Mallows Distribution



Topic Permutation

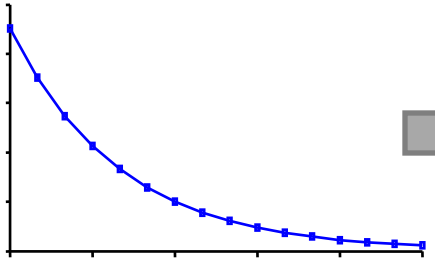


Topic Count Distribution

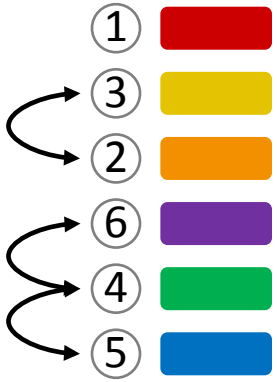
1	**3**
2	**2**
3	**1**
4	**3**
5	**1**
6	**2**

Topic Counts

Generative Process

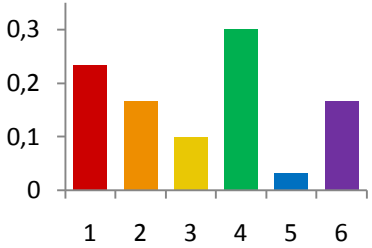


Mallows Distribution

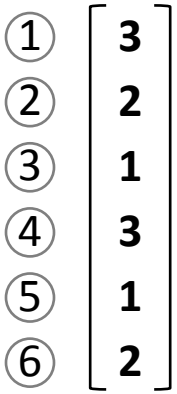


Topic Permutation

Document Topic Assignments

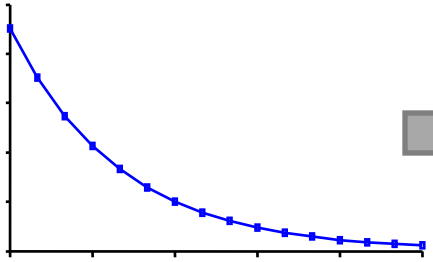


Topic Count Distribution

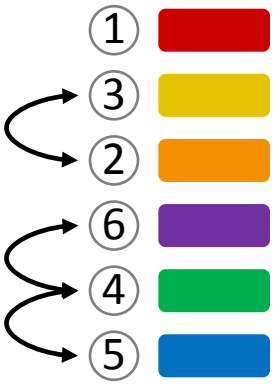


Topic Counts

Generative Process

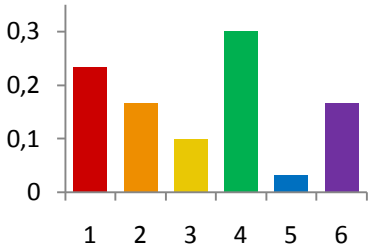


Mallows Distribution

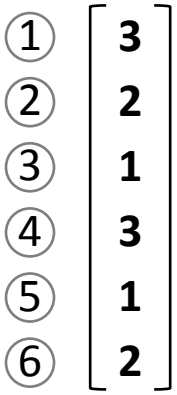


Topic Permutation

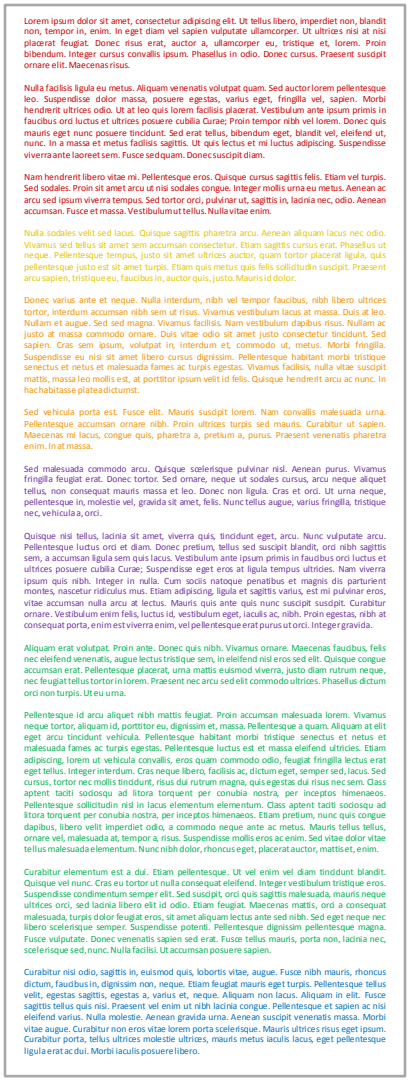
Document Topic Assignments



Topic Count Distribution



Topic Counts



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut tellus libero, imperdiet non, blandit non, tempor in, enim. In eget diam vel sapien vulputate ullamcorper. Ut ultrices nisi at nisi placerat feugiat. Donec risus erat, auctor a, ullamcorper eu, tristique et, lorem. Proin bibendum. Integer cursus convallis ipsum. Phasellus in odio. Donec cursus. Praesent suscipit ornare elit. Maecenas risus.

Nulla facilisis ligula eu metus. Aliquam venenatis volutpat quam. Sed auctor lorem pellentesque leo. Suspendisse dolor massa, posuere egestas, varius eget, fringilla vel, sapien. Morbi hendrerit ultrices odio. Ut at leo quis lorem facilisis placerat. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin tempor nibh vel lorem. Donec quis mauris eget nunc posuere tincidunt. Sed erat tellus, bibendum eget, blandit vel, eleifend ut, nunc. In a massa et metus facilisis sagittis. Ut quis lectus et mi luctus adipiscing. Suspendisse viverra ante laoreet sem. Fusce sed quam. Donec suscipit diam.

Nam hendrerit libero vitae mi. Pellentesque eros. Quisque cursus sagittis felis. Etiam vel turpis. Sed sodales. Proin sit amet arcu ut nisi sodales congue. Integer mollis urna eu metus. Aenean ac arcu sed ipsum viverra tempus. Sed tortor orci, pulvinar ut, sagittis in, lacinia nec, odio. Aenean accumsan. Fusce et massa. Vestibulum ut tellus. Nulla vitae enim.

Nulla sodales velit sed lacus. Quisque sagittis pharetra arcu. Aenean aliquam lacus nec odio. Vivamus sed tellus sit amet sem accumsan consectetur. Etiam sagittis cursus erat. Phasellus ut neque. Pellentesque tempus, justo sit amet ultrices auctor, quam tortor placerat ligula, quis pellentesque justo est sit amet turpis. Etiam quis metus quis felis sollicitudin suscipit. Praesent arcu sapien, tristique eu, faucibus in, auctor quis, justo. Mauris id dolor.

Donec varius ante et neque. Nulla interdum, nibh vel tempor faucibus, nibh libero ultrices tortor, interdum accumsan nibh sem ut risus. Vivamus vestibulum lacus at massa. Duis at leo. Nullam et augue. Sed sed magna. Vivamus facilisis. Nam vestibulum dapibus risus. Nullam ac justo at massa commodo ornare. Duis vitae odio sit amet justo consectetur tincidunt. Sed sapien. Cras sem ipsum, volutpat in, interdum et, commodo ut, metus. Morbi fringilla. Suspendisse eu nisi sit amet libero cursus dignissim. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vivamus facilisis, nulla vitae suscipit mattis, massa leo mollis est, at porttitor ipsum velit id felis. Quisque hendrerit arcu ac nunc. In hachabitasse platea dictumst.

Sed vehicula porta est. Fusce elit. Mauris suscipit lorem. Nam convallis malesuada urna. Pellentesque accumsan ornare nibh. Proin ultrices turpis sed mauris. Curabitur ut sapien. Maecenas mi lacus, congue quis, pharetra a, pretium a, purus. Praesent venenatis pharetra enim. In at massa.

Sed malesuada commodo arcu. Quisque scelerisque pulvinar nisl. Aenean purus. Vivamus fringilla feugiat erat. Donec tortor. Sed ornare, neque ut sodales cursus, arcu neque aliquet tellus, non consequat mauris massa et leo. Donec non ligula. Cras et orci. Ut urna neque, pellentesque in, molestie vel, gravida sit amet, felis. Nunc tellus augue, varius fringilla, tristique nec, vehicula a, orci.

Quisque nisi tellus, lacinia sit amet, viverra quis, tincidunt eget, arcu. Nunc vulputate arcu. Pellentesque luctus orci et diam. Donec pretium, tellus sed suscipit blandit, orci nibh sagittis sem, a accumsan ligula sem quis lacus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Suspendisse eget eros at ligula tempus ultricies. Nam viverra ipsum quis nibh. Integer in nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Etiam adipiscing, ligula et sagittis varius, est mi pulvinar eros, vitae accumsan nulla arcu at lectus. Mauris quis ante quis nunc suscipit suscipit. Curabitur ornare. Vestibulum enim felis, luctus id, vestibulum eget, iaculis ac, nibh. Proin egestas, nibh at consequat porta, enim est viverra enim, vel pellentesque erat purus ut orci. Integer gravida.

Aliquam erat volutpat. Proin ante. Donec quis nibh. Vivamus ornare. Maecenas faucibus, felis nec eleifend venenatis, augue lectus tristique sem, in eleifend nisl eros sed elit. Quisque congue accumsan erat. Pellentesque placerat, urna mattis euismod viverra, justo diam rutrum neque, nec feugiat tellus tortor in lorem. Praesent nec arcu sed elit commodo ultrices. Phasellus dictum orci non turpis. Ut eu uma.

Pellentesque id arcu aliquet nibh mattis feugiat. Proin accumsan malesuada lorem. Vivamus neque tortor, aliquam id, porttitor eu, dignissim et, massa. Pellentesque a quam. Aliquam at elit eget arcu tincidunt vehicula. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Pellentesque luctus est et massa eleifend ultricies. Etiam adipiscing, lorem ut vehicula convallis, eros quam commodo odio, feugiat fringilla lectus erat eget tellus. Integer interdum. Cras neque libero, facilisis ac, dictum eget, semper sed, lacus. Sed cursus, tortor nec mollis tincidunt, risus dui rutrum magna, quis egestas dui risus nec sem. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Pellentesque sollicitudin nisl in lacus elementum elementum. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Etiam pretium, nunc quis congue dapibus, libero velit imperdiet odio, a commodo neque ante ac metus. Mauris tellus tellus, ornare vel, malesuada at, tempor a, risus. Suspendisse mollis eros ac enim. Sed vitae dolor vitae tellus malesuada elementum. Nunc nibh dolor, rhoncus eget, placerat auctor, mattiset, enim.

Curabitur elementum est a dui. Etiam pellentesque. Ut vel enim vel diam tincidunt blandit. Quisque vel nunc. Cras eu tortor ut nulla consequat eleifend. Integer vestibulum tristique eros. Suspendisse condimentum semper elit. Sed suscipit, orci quis sagittis malesuada, mauris neque ultrices orci, sed lacinia libero elit id odio. Etiam feugiat. Maecenas mattis, ord a consequat malesuada, turpis dolor feugiat eros, sit amet aliquam lectus ante sed nibh. Sed eget neque nec libero scelerisque semper. Suspendisse potenti. Pellentesque dignissim pellentesque magna. Fusce vulputate. Donec venenatis sapien sed erat. Fusce tellus mauris, porta non, lacinia nec, scelerisque sed, nunc. Nulla facilisi. Ut accumsan posuere sapien.

Curabitur nisi odio, sagittis in, euismod quis, lobortis vitae, augue. Fusce nibh mauris, rhoncus dictum, faucibus in, dignissim non, neque. Etiam feugiat mauris eget turpis. Pellentesque tellus velit, egestas sagittis, egestas a, varius et, neque. Aliquam non lacus. Aliquam in elit. Fusce sagittis tellus quis nisi. Praesent vel enim ut nibh lacinia congue. Pellentesque et sapien ac nisi eleifend varius. Nulla molestie. Aenean gravida urna. Aenean suscipit venenatis massa. Morbi vitae augue. Curabitur non eros vitae lorem porta scelerisque. Mauris ultrices risus eget ipsum. Curabitur porta, tellus ultrices molestie ultrices, mauris metus iaculis lacus, eget pellentesque ligula erat ac dui. Morbi iaculis posuere libero.

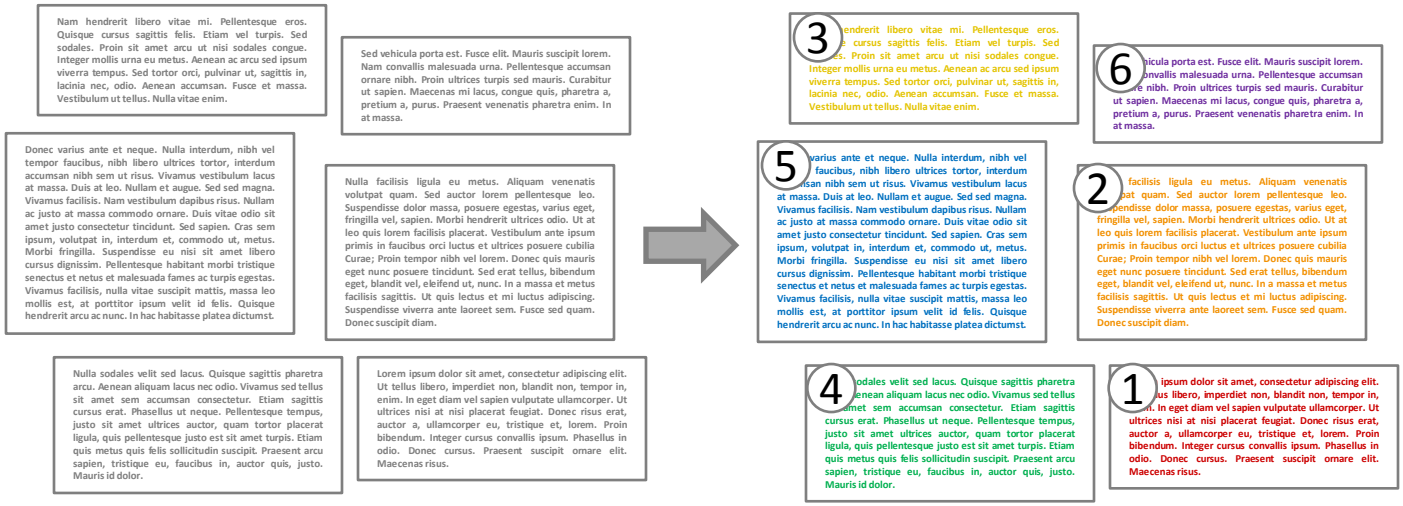
Training Details

- Hidden variables:
  - *Predominant ordering, Mallows parameters*
  - *Topic counts*, *language models*

- Estimated from unannotated documents using Bayesian inference techniques
  - *Collapsed Gibbs sampling*

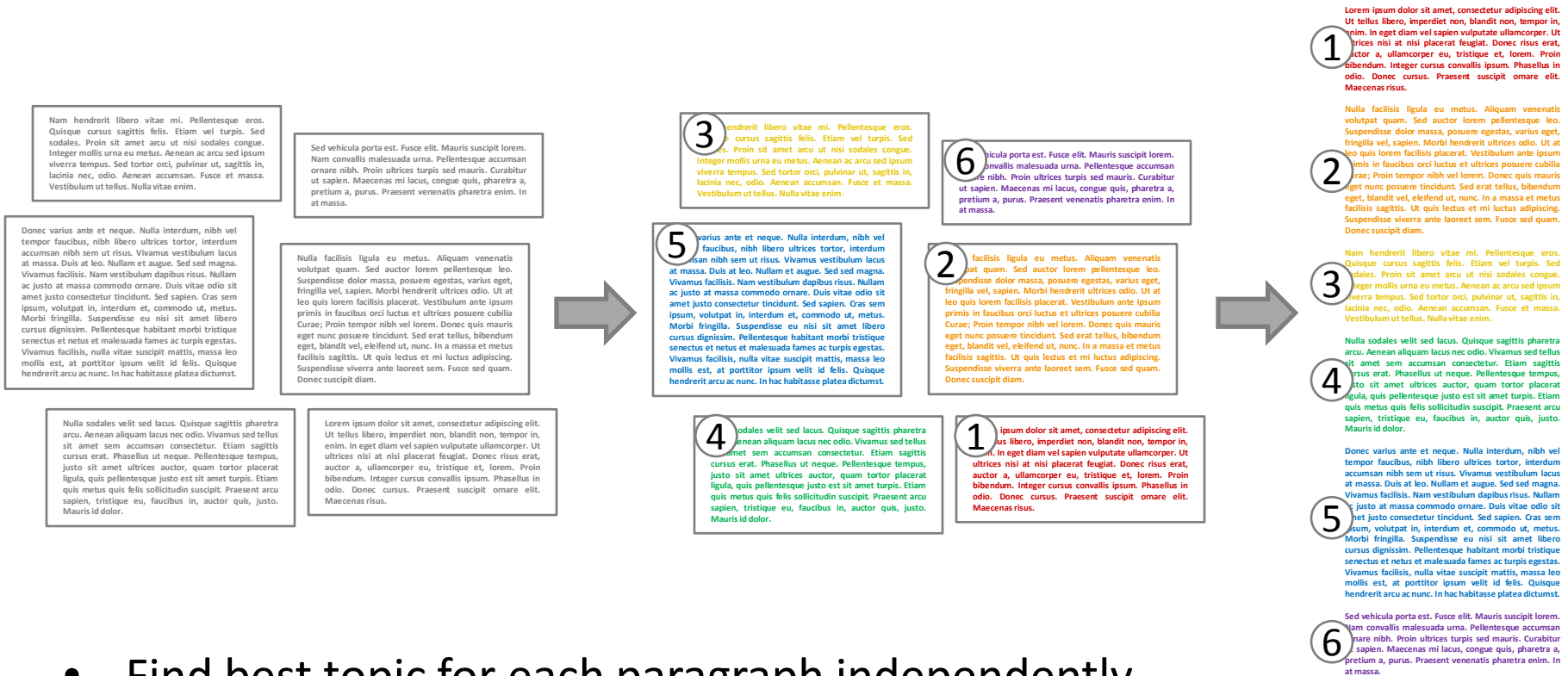
- Mallows distribution admits computationally tractable inference procedures

Predicting Document Ordering



- Find best topic for each paragraph independently
  - Computationally simple, because marginalization over all orderings is not necessary!

Predicting Document Ordering



- Find best topic for each paragraph independently
  - Computationally simple, because marginalization over all orderings is not necessary!
- Sort paragraphs according to topic number
  - Topic assignments determine best ordering

Evaluation Results: Ordering

Average number of sentences per document: 10

Model	Disasters	NTSB Reports
Content Models (HMM)	81.0	44.0
Content Models (GMM)	83.0	57.0
Random	<1.0	<1.0

See NAACL'09 for evaluation on segmentation and semantic labeling

Outline

- Applications of Text Models
- Coherence Models (ACL 2005, CL 2008)
- Content Models (NAACL 2004, NAACL 2009)

Modeling Coherence

Active networks and virtual machines have a long history of collaborating in this manner. The basic tenet of this solution is the refinement of Scheme. The disadvantage of this type of approach, however, is that public-private key pair and red-black trees are rarely incompatible.

- **Coherence** is a property of well-written texts that makes them easier to read and understand than a sequence of randomly strung sentences
- **Local coherence** captures text organization at the level of sentence-to-sentence transitions

Our Approach

Key Premise: the distribution of entities in locally coherent discourse exhibits certain regularities

- Abstract a text into an entity-based representation that encodes syntactic and distributional information
- Learn properties of coherent texts, given a training set of coherent and incoherent texts

Our Inspiration: Centering Theory

Grosz&Joshi&Weinstein,1983; Strube&Hahn,1999; Poesio&Stevenson&Di Eugenio&Hitzeman,2004

- Constraints on the entity distribution in a coherent text
  - Focus is the most salient entity in a discourse segment
  - Transition between adjacent sentences is characterized in terms of focus switch
- Constraints on linguistic realization of focus
  - Focus is more likely to be realized as subject or object
  - Focus is more likely to be referred to with anaphoric expression

Text Representation

- Entity Grid — a two-dimensional array that captures the distribution of discourse entities across text sentences
- Discourse Entity — a class of coreferent noun phrases

Input Text

1 Former Chilean dictator Augusto Pinochet, was arrested in London on October 14th, 1998.

2 Pinochet, 82, was recovering from surgery.

3 The arrest was in response to an extradition warrant served by a Spanish judge.

4 Pinochet was charged with murdering thousands, including many Spaniards.

5 He is awaiting a hearing, his fate in the balance.

6 American scholars applauded the arrest.

Input Text with Syntactic Annotation

Use Collins' parser(1997):

1. [Former Chilean dictator Augusto Pinochet]$_{\mathbf{S}}$, was arrested in [London]$_{\mathbf{X}}$ on [October 14th]$_{\mathbf{X}}$ 1998.
2. [Pinochet]$_{\mathbf{S}}$, 82, was recovering from [surgery]$_{\mathbf{X}}$.
3. [The arrest]$_{\mathbf{S}}$ was in [response]$_{\mathbf{X}}$ to [an extradition warrant]$_{\mathbf{X}}$ served by [a Spanish judge]$_{\mathbf{S}}$.
4. [Pinochet]$_{\mathbf{S}}$ was charged with murdering [thousands]$_{\mathbf{O}}$, including many [Spaniards]$_{\mathbf{O}}$.
5. [He]$_{\mathbf{S}}$ is awaiting [a hearing]$_{\mathbf{O}}$, [his fate]$_{\mathbf{X}}$ in [the balance]$_{\mathbf{X}}$.
6. [American scholars]$_{\mathbf{S}}$ applauded the [arrest]$_{\mathbf{O}}$.

Notation: **S**=subjects, **O**=object, **X**=other

Input Text with Coreference Information

Use noun-phrase coreference tool (Ng and Cardie, 2002):

1. [Former Chilean dictator Augusto Pinochet]$_{\mathbf{S}}$, was arrested in [London]$_{\mathbf{X}}$ on [October 14]$_{\mathbf{X}}$ 1998.
2. [Pinochet]$_{\mathbf{S}}$, 82, was recovering from [surgery]$_{\mathbf{X}}$.
3. [The arrest]$_{\mathbf{S}}$ was in [response]$_{\mathbf{X}}$ to [an extradition warrant]$_{\mathbf{X}}$ served by [a Spanish judge]$_{\mathbf{S}}$.
4. [Pinochet]$_{\mathbf{S}}$ was charged with murdering [thousands]$_{\mathbf{O}}$, including many [Spaniards]$_{\mathbf{O}}$.
5. [He]$_{\mathbf{S}}$ is awaiting [a hearing]$_{\mathbf{O}}$, [his fate]$_{\mathbf{X}}$ in [the balance]$_{\mathbf{X}}$.
6. [American scholars]$_{\mathbf{S}}$ applauded the [arrest]$_{\mathbf{O}}$.

Output Entity Grid

	Pinochet	London	October	Surgery	Arrest	Extradition	Warrant	Judge	Thousands	Spaniards	Hearing	Fate	Balance	Scholars	
1	s	x	x	–	–	–	–	–	–	–	–	–	–	–	1
2	s	–	–	x	–	–	–	–	–	–	–	–	–	–	2
3	–	–	–	–	s	x	x	s	–	–	–	–	–	–	3
4	s	–	–	–	–	–	–	–	o	o	–	–	–	–	4
5	s	–	–	–	–	–	–	–	–	–	o	x	x	–	5
6	–	–	–	–	o	–	–	–	–	–	–	–	–	s	6

Comparing Grids

S	S	S	X	X	–	–	–	–	–	–	–	–	–	–
–	–	S	–	–	X	–	–	–	–	–	–	–	–	–
–	–	–	–	–	–	S	X	X	O	–	–	–	–	–
–	–	S	–	–	–	–	–	–	–	O	O	–	–	–
–	–	S	–	–	–	–	–	–	–	–	–	O	X	X
–	–	–	–	–	–	O	–	–	–	–	–	–	–	–

S	S	X	X	X	–	–	–	–	–	–	–	–	–	X
–	–	X	–	–	X	–	–	–	–	–	–	–	–	X
–	–	X	–	–	–	–	X	X	O	–	–	–	–	X
–	–	X	–	–	–	–	–	–	–	O	O	–	–	X
–	–	X	–	–	–	–	–	–	–	–	–	O	X	X
–	–	X	–	–	–	O	–	–	–	–	–	–	–	X

Text Encoding as Feature Vector

	S S	S O	S X	S |	O S	O O	O X	O |	X S	X O	X X	X |	| S	| O	| X	| |
d_{i1}	0	0	0	.03	0	0	0	.02	.07	0	0	.12	.02	.02	.05	.25
d_{i2}	.02	0	0	.03	0	0	0	.06	0	0	0	.05	.03	.07	.07	.29

Each grid rendering x_{ij} of a document d_i is represented by a feature vector:

$$\Phi(x_{ij}) = (p_1(x_{ij}), p_2(x_{ij}), \ldots, p_m(x_{ij}))$$

where m is the number of all predefined entity transitions, and $p_t(x_{ij})$ the probability of transition t in the grid x_{ij}

Learning a Ranking Function

- **Training Set**

  Ordered pairs (x_{ij}, x_{ik}), where x_{ij} and x_{ik} are renderings of the same document d_i, and x_{ij} exhibits a higher degree of coherence than x_{ik}

- **Training Procedure**
  - Goal: Find a parameter vector $\vec{w}$ that yields a "ranking score" function $\vec{w} \cdot \Phi(x_{ij})$ satisfying:

$$\vec{w} \cdot (\Phi(x_{ij}) - \Phi(x_{ik})) > 0$$
$$\forall (x_{ij}, x_{ik}) \text{ in training set}$$

  - Method: Constraint optimization problem solved using the search technique described in Joachims (2002)

Evaluation: Information Ordering

- Goal: recover the most coherent sentence ordering
- Basic set-up:
  - Input: a pair of a source document and a permutation of its sentences

  - Task: find a source document via coherence ranking
- Data: Training 4000 pairs, Testing 4000 pairs (Natural disasters and Transportation Safety Reports)

Evaluation: Summarization

- Goal: select the most coherent summary among several alternatives
- Basic set-up:
  - Input: a pair of system summaries
  - Task: predict the ranking provided by human
- Data: 96 summary pairs for training, 32 pairs for testing (from DUC 2003)

Evaluation Results

Model	Disasters	NTSB reports	Summary Ranking
Grid	87.2	90.4	80.0
Content Model (HMM)	88.0	75.8	N/A
Random	50	50	50

Follow-Ups

- Elsner, Austerweil, and Charniak (NAACL 2007): combining content and coherence models is effective
  - **Brown Coherence Toolkit**: software for coherence models and test applications

- Pitler and Nenkova (EMNLP 2008): grid-based features help to predict readability of human-authored texts

Conclusions

- Practical Benefits: The proposed models can be easily incorporated in text-to-text generation
- Main Findings: The proposed document models can be successfully applied to modeling text structure
- Future Research: Design statistical models that match representational power of traditional discourse theories

References

- Harr Chen, S.R.K. Branavan, Regina Barzilay, David R. Karger. “*Latent Topic Models for Document Structure Induction*”, *NAACL-HLT, 2009.*
- Regina Barzilay, Mirella Lapata *"Modeling Local Coherence: An Entity-based Approach"*, Computational Linguistics, 2008.
- Regina Barzilay, Lillian Lee *"Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization"*, NAACL-HLT, 2004