Text Segmentation

Regina Barzilay

MIT

October, 2005

Linear Discourse Structure: Example

Stargazers Text(from Hearst, 1994)

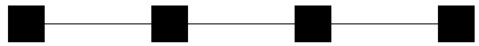
- Intro - the search for life in space
- The moon's chemical composition
- How early proximity of the moon shaped it
- How the moon helped life evolve on earth
- Improbability of the earth-moon system

What is Segmentation?

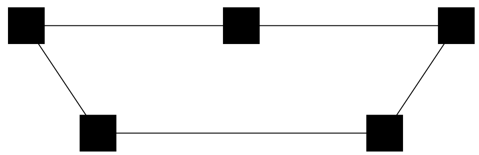
Segmentation: determining the positions at which topics change in a stream of text or speech.

SEGMENT 1: OKAY
tsk There's a farmer,
he looks like ay uh Chicano American,
he is picking pears.
A-nd u-m he's just picking them,
he comes off the ladder,
a-nd he- u-h puts his pears into the basket.
SEGMENT 2: U-h a number of people are going by,
and one of them is um I don't know,
I can't remember the first . . . the first person that goes by

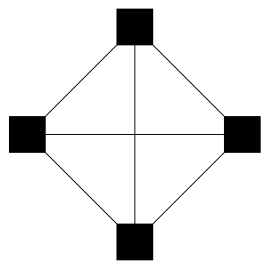
Skorochodko's Text Types



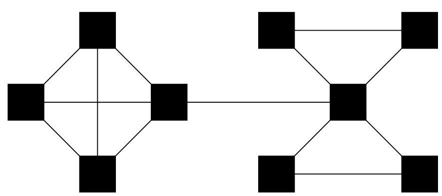
Chained



Ringed



Monolith

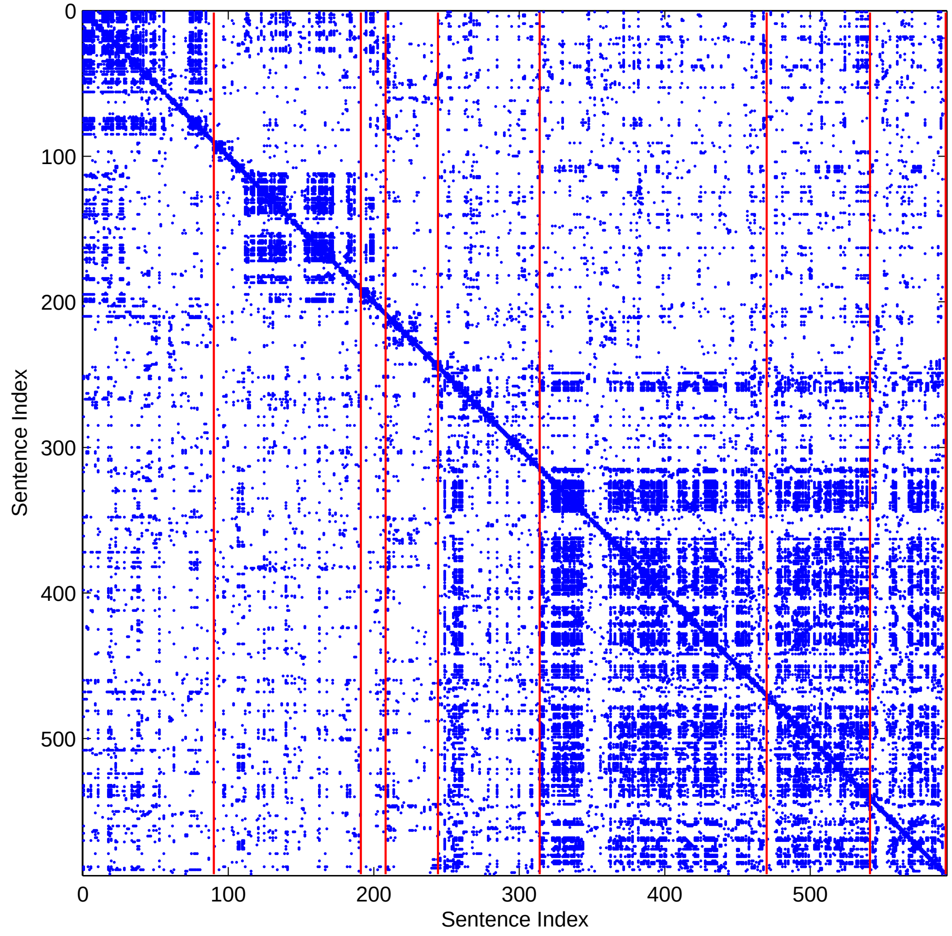


Piecewise

Word Distribution in Text

Table removed for copyright reasons.

Please see: Figure 2 in Hearst, M. "Multi-Paragraph Segmentation of Expository Text." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 94)*, June 1994. (http://www.sims.berkeley.edu/~hearst/papers/tiling-acl94/acl94.html)



0
100
200
300
400
500
Sentence Index
Sentence Index

Today

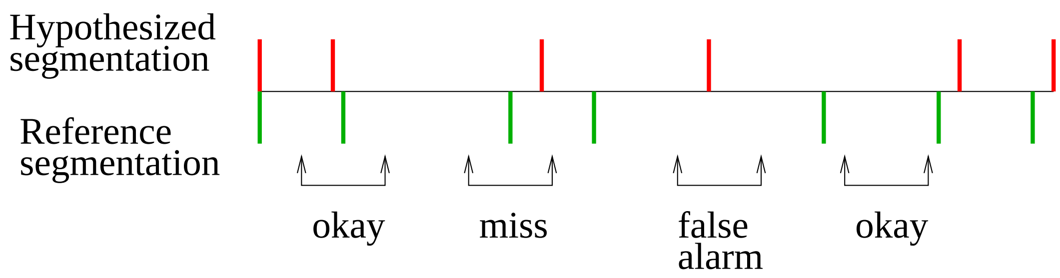
- Evaluation measures
- Similarity-based segmentation
- Feature-based segmentation

Evaluation Measures

- Precision (P): the percentage of proposed boundaries that exactly match boundaries in the reference segmentation
- Recall (R): the percentage of reference segmentation boundaries that are proposed by the algorithm
- $F = 2\frac{PR}{(P+R)}$

Problems?

Evaluation Metric: P_k Measure



P_k: Probability that a randomly chosen pair of words k words apart is inconsistently classified (Beeferman '99)

- Set k to half of average segment length
- At each location, determine whether the two ends of the probe are in the same or different location. Increase a counter if the algorithm's segmentation disagree with the reference segmentation

- Normalize the count between 0 and 1 based on the number of measurements taken

Notes on P_k measure

- $P_k \in [0, 1]$, the lower the better
- Random segmentation: $P_k \approx 0.5$
- On synthetic corpus: $P_k \in [0.05, 0.2]$
- Beeferman reports 0.19 P_k on WSJ, 0.13 on Broadcast News

Corpus

- Synthetic data
  - Choi'2000: concatenate paragraphs from different texts
- Broadcast news (stories are not segmented)
- Manually segmented material (texts, lectures, meetings)

Cohesion

Key hypothesis: cohesion ties reflect text structure
Cohesion captures devices that link sentences into a text (Halliday&Hasan)

- Lexical cohesion
- References
- Ellipsis
- Conjunctions

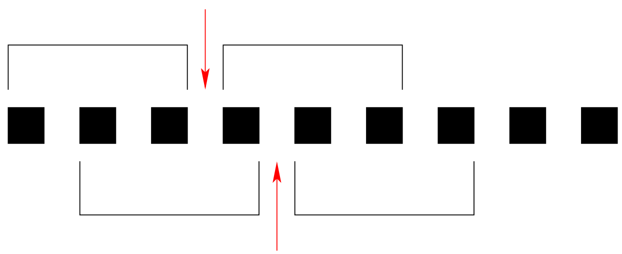
Word Distribution in Text

Table removed for copyright reasons.

Please see: Figure 2 in Hearst, M. "Multi-Paragraph Segmentation of Expository Text." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 94)*, June 1994. (http://www.sims.berkeley.edu/~hearst/papers/tiling-acl94/acl94.html)

Segmentation Algorithm of Hearst

- Preprocessing and Initial segmentation
- Similarity Computation
- Boundary Detection



Similarity Computation: Representation

Vector-Space Representation

SENTENCE$_1$: I like apples

SENTENCE$_2$: Apples are good for you

Vocabulary	Apples	Are	For	Good	I	Like	you
Sentence$_1$	1	0	0	0	1	1	0
Sentence$_2$	1	1	1	1	0	0	1

Similarity Computation: Cosine Measure

Cosine of angle between two vectors in n-dimensional space

$$sim(b_1, b_2) = \frac{\sum_t w_{y,b_1} w_{t,b_2}}{\sqrt{\sum_t w_{t,b_1}^2 \sum_{t=1}^n w_{t,b_2}^2}}$$

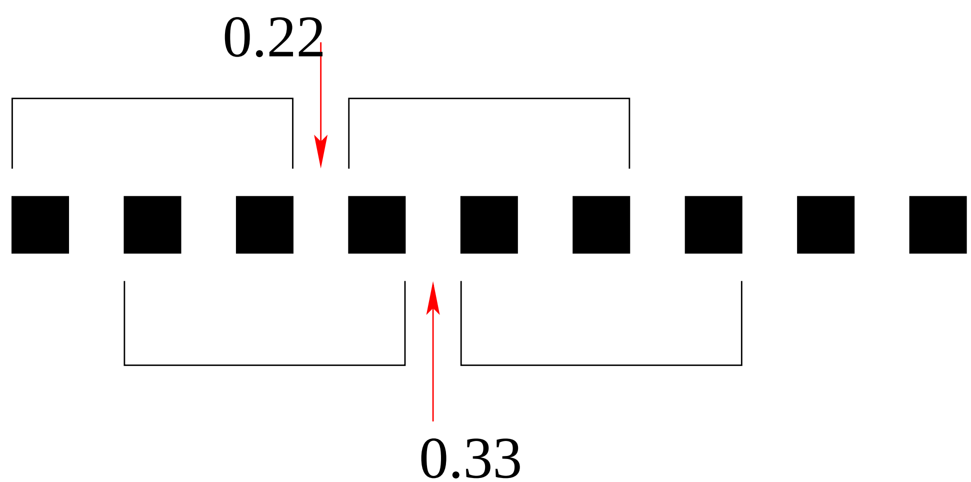
$SENTENCE_1$: 1 0 0 0 1 1 0

$SENTENCE_2$: 1 1 1 1 0 0 1

$sim(S_1,S_2)$ =

$$\frac{1*0+0*1+0*1+0*1+1*0+1*0+0*1}{\sqrt{(1^2+0^2+0^2+0^2+1^2+1^2+0^2)*(1^2+1^2+1^2+1^2+0^2+0^2+1^2}} = 0.26$$

Similarity Computation: Output



Gap Plot

Figure of Gap Plot removed for copyright reasons.

Boundary Detection

Boundary detection is based on changes in sequence of similarity scores

- Compute depth scores for each gap i
  - Find closest maximum on the left and subtract it from i
  - Find closest maximum on the right and subtract it from i
  - Sum right and left scores
- Sort depth scores and select k boundaries

  - Number of segments is determined by the depth score threshold: $s - \sigma/2$

- Incorporate constraints on sentence length and adjust for paragraph breaks

Segmentation Evaluation

Comparison with human-annotated segments(Hearst'94):

- 13 articles (1800 and 2500 words)
- 7 judges
- boundary if three judges agree on the same segmentation point

Agreement on Segmentation

Figure removed for copyright reasons.

Please see: Figure 3 in Hearst, M. "Multi-Paragraph Segmentation of Expository Text." *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 94)*, June 1994. (http://www.sims.berkeley.edu/~hearst/papers/tiling-acl94/acl94.html)

Evaluation Results

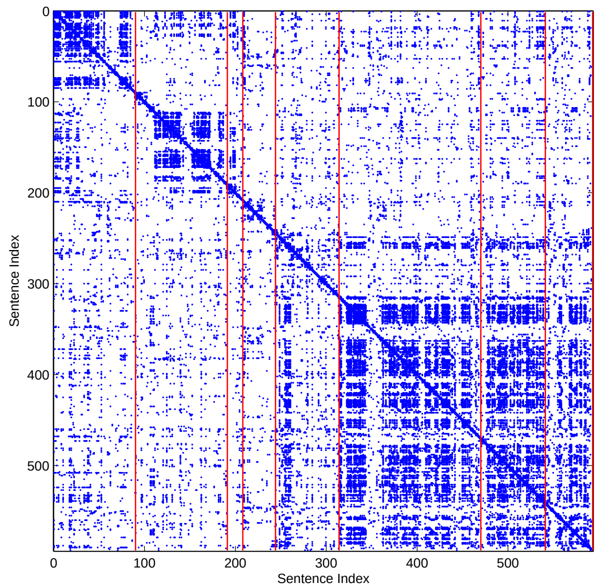
Methods	Precision	Recall
Baseline 33%	0.44	0.37
Baseline 41%	0.43	0.42
Chains	0.64	0.58
Blocks	0.66	0.61
Judges	0.81	0.71

More Results

- High sensitivity to change in parameter values
- Thesaural information does not help
- Most of the mistakes are “close misses”

Our Approach: Min Cut Segmentation

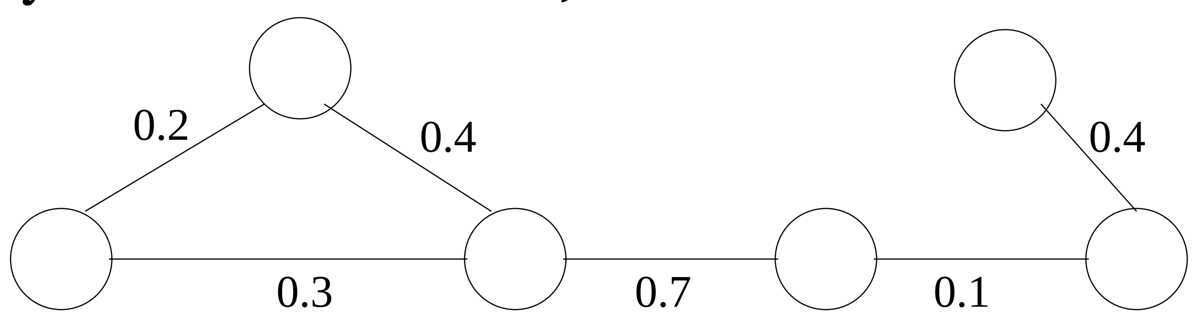
- Key assumption: change in lexical distribution signals topic change (Hearst '94)



- Goal: identify regions of lexical cohesiveness
  - Method: Min Cut Graph Partitioning

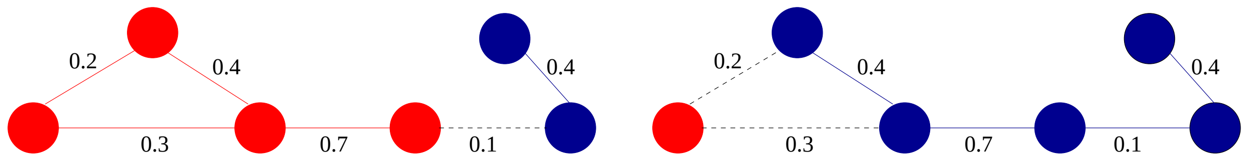
Graph-based Representation

- Let $G(V, E)$ be a weighted undirected graph
  - V - set of nodes in the graph
  - E - set of weighted edges
- Edge weights $w(u, v)$ define a measure of pairwise similarity between nodes u,v



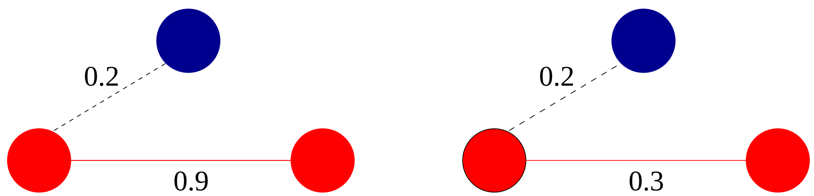
Definitions

- Graph cut: partitioning of the graph into two disjoint sets of nodes A,B
- Graph cut weight: $\text{cut}(A, B) = \sum_{u \in A, v \in B} w(u, v)$
  – i.e. sum of crossing edge weights
- Minimum Cut: the cut that minimizes cross-partition similarity



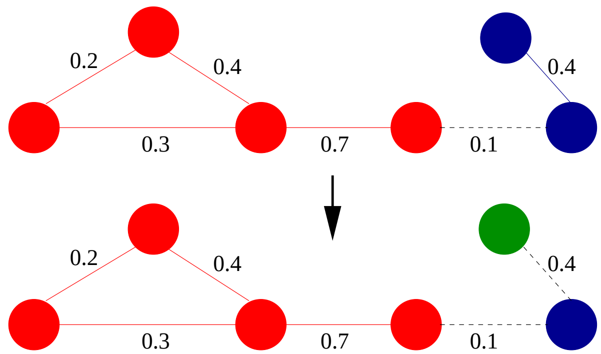
Normalized Cuts

- Motivation: need to account for intra-partition similarity



- Shi and Malik, 1999: normalize the cut by the partition volume
  - Volume is the total weight of edges from the set to other nodes in G
  - vol(A) = $\sum_{u \in A, v \in V} w(u, v)$
- $Ncut(A, B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}$

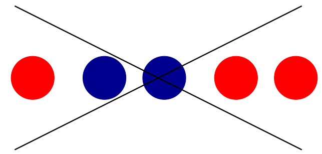
Multi-partitioning problem



- $Ncut_k(V) = \frac{cut(A_1, A_1 - V)}{vol(A_1)} + \ldots + \frac{cut(A_k, A_k - V)}{vol(A_k)}$
  – where $A_1, \ldots A_k$ are the partitioning sets
- Multi-way partitioning problem is NP-complete (Papadimitrious, '99)

Computing the optimal Multi-Way Partitioning

- Partitions need to preserve linearity of segmentation



- Exact solution can be found using dynamic programming in polynomial time

$$\min Ncut_k(V) = \min_{A \subset V} Ncut_{k-1}(V - A) + \frac{cut(A, V - A)}{vol(A)}$$

Text Segmentation with Minimum Cuts

- Sentences are represented by nodes in the graph
- Graph is fully connected
  - Edge weights computed between every pair of nodes
- Weight of an edge (s_i, s_j): $w(s_i, s_j) = e^{\frac{s_i \cdot s_j}{||s_i|| \times ||s_j||}}$

Additional Model Parameters

- Granularity:
  - Fixed window size vs sentence representation
- Lexical representation:
  - Stop words removal
  - Word stemming with Porter's stemmer
  - Technical term extraction
- Similarity Computation:
  - Word Occurrence smoothing:
    $\tilde{s}_{i+k} = \sum_{j=i}^{i+k} e^{-\alpha(i+k-j)} s_j$

Experimental Results

Algorithm	AI	Physics
Random	0.49	0.5
Uniform	0.52	0.46
MinCutSeg	0.35	0.36

Human Evaluation

Lecture Id	Annotator	P_k Measure
1	A	0.23
1	B	0.23
2	A	0.37
2	B	0.36
P_k Average		0.3

Advantages of Min Cut Segmentation

- Unsupervised learning method
- Supports global inference
- Computes efficiently

Simple Feature-Based Segmentation

Litman&Passanneau'94

- Prosodic Features:
  - pause: true, false
  - duration: continuous
- Cue Phrase Features:
  - Word: *also, and, anyway, basically, because, oh, okay, see, so, well*

Results

	Recall	Precision	Error
Cue	72%	15%	50%
Pause	92%	18%	49%
Humans	74%	55%	11%

Possible Features

- Does the word appear up to 1 sentence in the future? 2 sentences? 3? 5?
- Does the word appear up to 1 sentence in the past? 2 sentences? 3? 5?
- Does the word appear up to 1 sentence in the past but not 5 sentences in the future?

Supervised Segmentation

- Goal: find a probability distribution $q(b|w)$, where $b \in \{YES, NO\}$ is a random variable describing the presence of a segment boundary in context w

- Desired distribution from the linear exponential family $Q(f, q_0)$ of the form:

$$Q(f, q_0) = \{q(b|w) = \frac{1}{Z_\lambda(w)} e^{\lambda \times f(w)} q_0(b|w)\},$$

$q_0(b|w)$ is a prior on the presence of the boundary
$\lambda \times f(w) = \lambda_1 \times f_1(w) + \ldots + \lambda_n \times f_n(w)$, where
$f_i(w) \in \{0, 1\}$
$Z_\lambda(w) = 1 + e^{\lambda \times f(w)}$ is a normalization constant

Supervised Segmentation

- Fitness function: KL divergence between $q \in Q(f, q_0)$ relative to a reference distribution of a sample of training events {(w,b)}

$$D(p||q) = \sum_{w} p(w) \sum_{b \in \{YES, NO\}} p(b|w) log \frac{p(b|w)}{q(b|w)}$$

- Parameter estimation method: iterative scaling

Results (WSJ)

	Recall	Precision	F
Model	54%	56%	55%
Random	16%	17%	17%
Even	17%	17%	17%