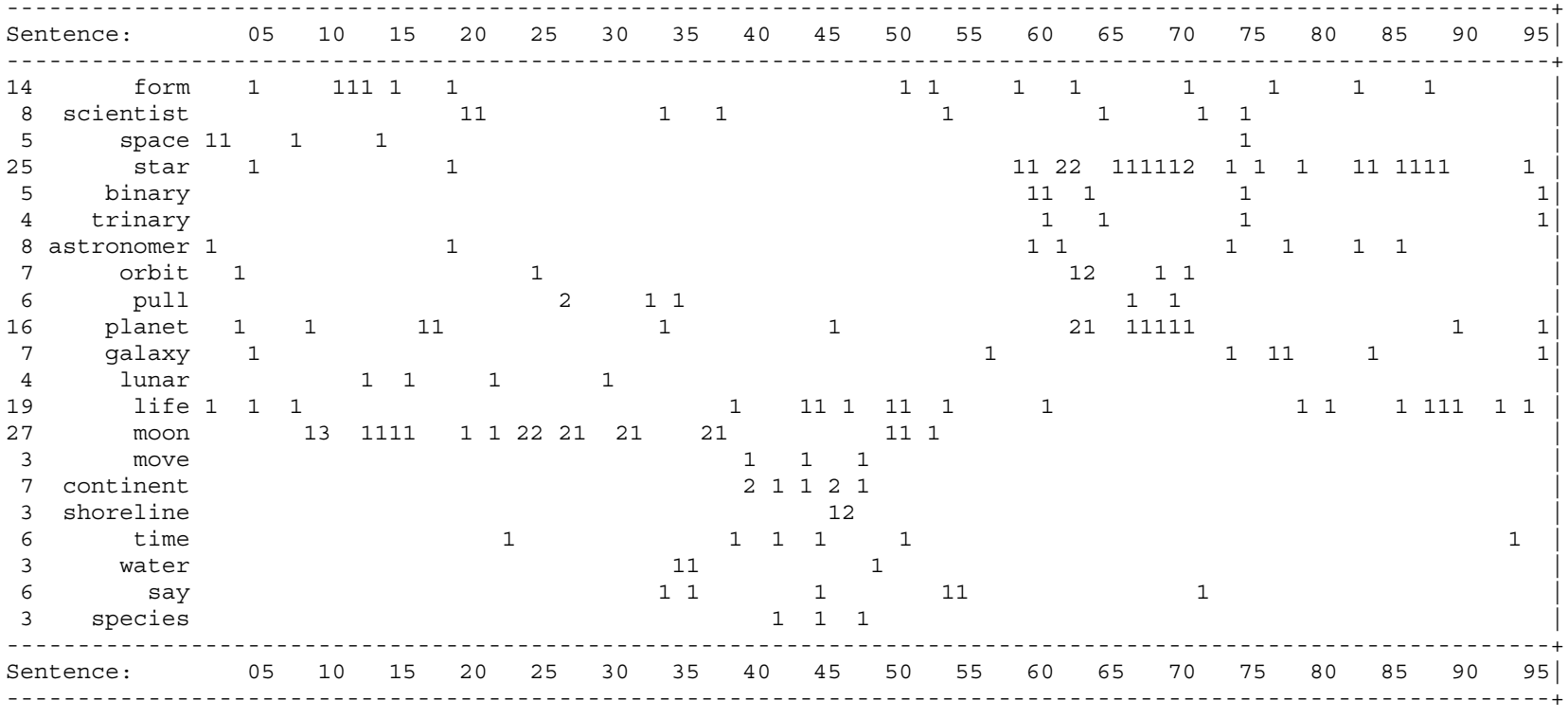
Topic Segmentation

Regina Barzilay

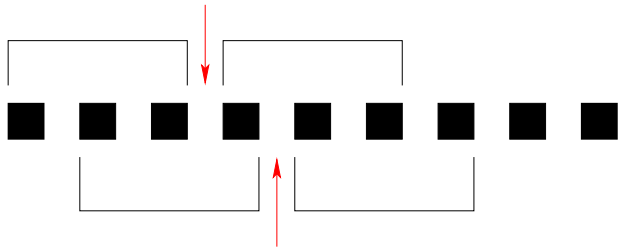
February 11, 2004

Example


Sentence: 05 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
14 form 1 111 1 1 1 1 1 1 1 1
8 scientist 11 1 1 1 1 1 1
5 space 11 1 1 1
25 star 1 1 1 11 22 111112 1 1 1 11 1111 1
5 binary 11 1 1 1
4 trinary 1 1 1 1
8 astronomer 1 1 1 1 1 1 1 1
7 orbit 1 1 12 1 1
6 pull 2 1 1 1 1
16 planet 1 1 11 1 1 21 11111 1 1
7 galaxy 1 1 1 11 1 1
4 lunar 1 1 1 1 1
19 life 1 1 1 1 11 1 11 1 1 1 1 1 111 1 1
27 moon 13 1111 1 1 22 21 21 21 11 1
3 move 1 1 1
7 continent 2 1 1 2 1
3 shoreline 12
6 time 1 1 1 1 1 1
3 water 11 1 1
6 say 1 1 1 11 1
3 species 1 1 1
Sentence: 05 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95

Segmentation Algorithm

- Preprocessing and Initial segmentation
- Similarity Computation
- Boundary Detection



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

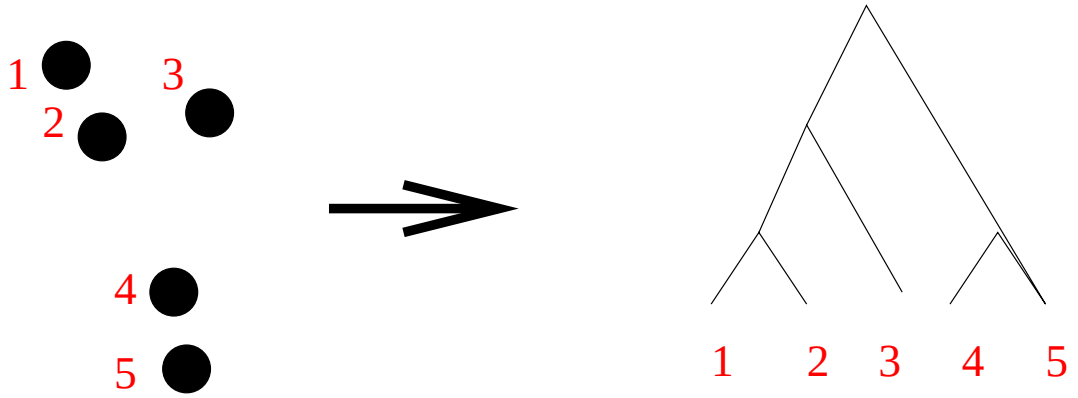
More Results

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

Today's Topics

- Hierarchical segmentation
- HMM-based segmentation
- Supervised segmentation

Agglomerative Clustering



- First, each data point is a singleton cluster
- Next, closest points are merged until all points are combined

Agglomerative Clustering

- Complete-link — merge the two clusters whose merger has the smallest diameter
- Single-link — merge the two clusters whose two closest members have the smallest distance
- Average-link — merges in each iteration the pair of clusters with the highest cohesion.

Hierarchical Segmentation

(Yaari, 1997)

- Partition the text into elementary segments
- While more than one segment left do
  - Find closest adjacent segments s_i, s_{i+1} (based on cosine measure)
  - Merge s_i, s_{i+1} into one segment

Broadcast News Segmentation

- Goal: divide news stream into stories
- Assumption: news stories typically belong to one of several categories (sports, politics, . . .)

HMM-based Segmentation: Construction

van Mulbregt&Carp&Gillick&Lowe'99:

- Each state of HMM represents a topic
- Topics are derived via story clustering
- Emission probabilities for a state are computed based on a unigram language model

HMM-based Segmentation: Decoding

- Transitions are controlled by switch penalty
- Segmentation via Viterbi-style decoding

TDT Segmentation Results

- Data: 384 shows, 6,000 stories and 2.2 million words
- Sources: ABC, CNN, ...
- TDT Evaluation Measure:
  $C_{Seg} = \alpha * P_{Miss} + (1 - \alpha) * P_{FalseAlarm}$

TDT Performance

Input Type	C_{Seg} for ABC
ASR	0.1723
Closed Captions	0.1515
Transcripts	0.1356

Note the impact for ASR!

Meeting Segmentation

- Motivation: Facilitate information Access
- Challenges:
  - High error rate in transcription
  - Multi-thread structure

Algorithm for Feature Segmentation

Supervised ML
(Galley&McKeown&Fosler-Lussier&Jing'03)

- Combines multiple knowledge source:
  - cue phrases
  - silences
  - overlaps
  - speaker change
  - lexical cohesion
- Uses probabilistic classifier (decision tree) to combine them

Cue Word Selection

Automatic computation of cue words:

- Compute word probability to appear in boundary position
- Select words with the highest probability
- Remove non-cues.

Selected Cue Words

OKAY	93.05
shall	0.44
anyway	0.43
alright	0.64
let's	0.66
good	0.81

Silences

- Pauses — speaker silence in the middle of her speech
- Gap — silences not attributable to any party

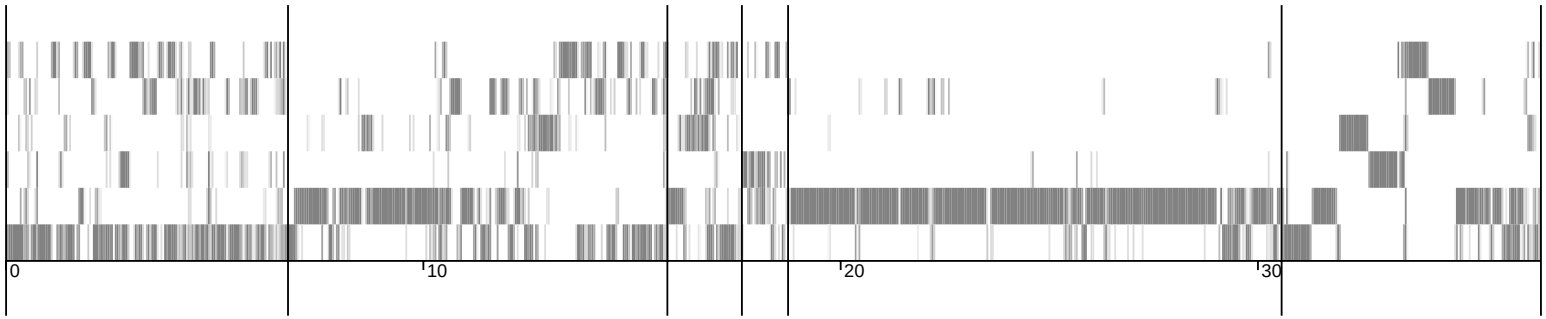
Topic boundaries are typically preceeded by gaps

Overlaps

- Average overlap rate within some window

Little overlap in the beginning of segments

Speaker Change



Determination of Window Size

Feature	Tag	Size(sec)	Side
Cue phrases	CUE	5	both
Silence (gaps)	SIL	30	left
Overlap	OVR	30	right
Speaker activity	ACT	5	both
Lexical cohesion	LC	30	both

Examples of Derived Rules

Condition	Decision	Conf.
LC$\leq$0.67, CUE$\geq$1, OVR$\leq$1.20, SIL$\leq$3.42	yes	94.1
LC$\leq$0.35, SIL$>$3.42, OVR$\leq$4.55	yes	92.2
CUE$\geq$1, ACT$>$0.1768, OVR$\leq$1.20, LC$\leq$0.67	yes	91.6
...		
default	no	

Results

Method	P_k	WD
Feature-based	23.00	25.47
Cohesion-based	31.91	35.88