MITRE TDT-2000 Segmentation System

Warren R. Greiff
Alex Morgan
Randall Fish
Marc Richards
Amlan Kundu

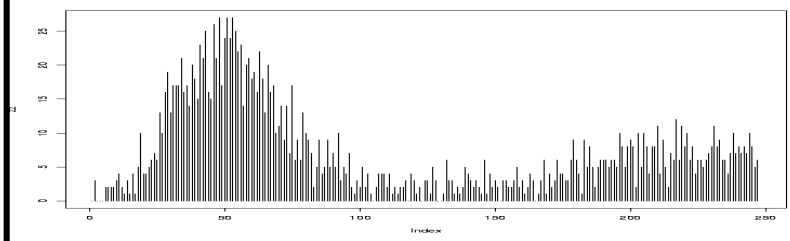
November, 2000

MITRE Segmentation Systems

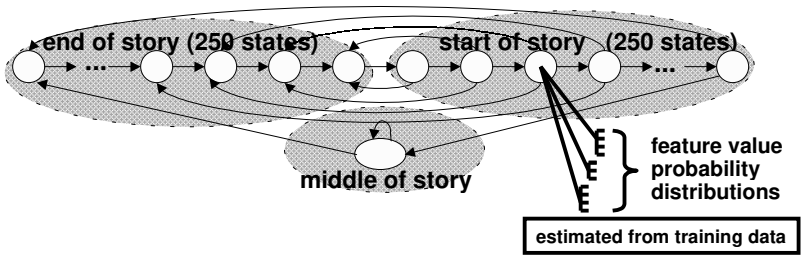
- **TDT-3**
  - **Naïve Bayes classifier**
  - **independent boundary/no-boundary decisions at each word**
  - **features:**
    - overlap
    - x-duration
    - triggers
    - audio cues
- **TDT-2000**
  - **fine-grained Hidden Markov Model**
    - story-length modeling
    - story-position features
    - single coherent segmentation

Story length profiles

- Each source has it's own typical profile
- ABC_WNT:

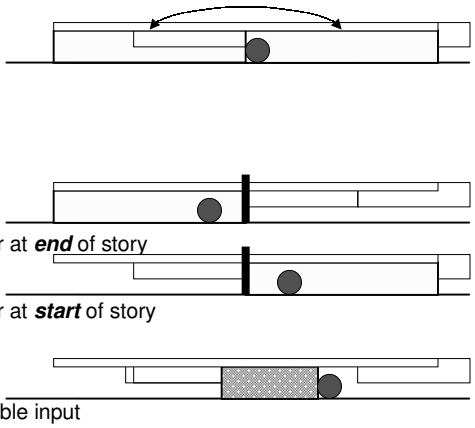


HMM: initial topology

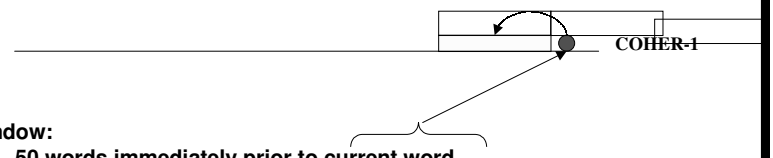


Features - TDT-3

- **overlap:**
  - amount of overlap
  - weighted by rarity
- **delta-overlap:**
  - change in overlap
- **pre-trigger cues:**
  - words that tend to occur at ***end*** of story
- **post-triggers:**
  - words that tend to occur at ***start*** of story
- **X-duration cue:**
  - duration of untranscribable input



TDT-2000: Coherence Features

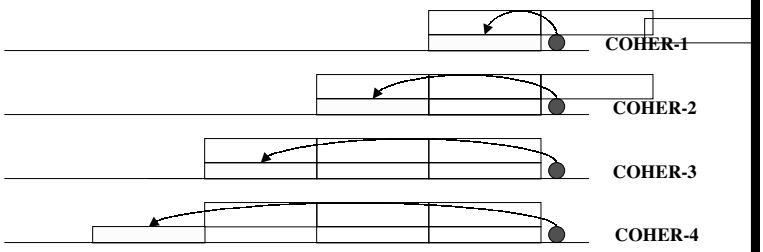


- **window:**
  50 words immediately prior to current word
- **value:**
  - **if not in window:** 0
  - **if occurs in window:** $-\log\frac{s_w}{s}$

s_w: # of stories in which w occurs
s: total # of stories

- **low values should be common at beginning of story**
- **high values more common 50 words or more into story**

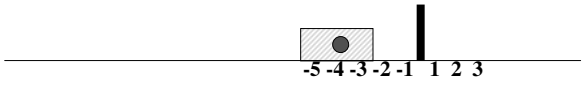
Features - coherence



- **motivation**
  - **indication of how deep we are into story**
  - **e.g.: high value of COHER-3:**
    - unlikely prior to word 100 of story
    - most likely after word 150 of story
    - increasing probability from 100 - 150

Trigger Features

- **Exploit:**
  - **words more likely to appear in specified regions (near boundary)**



  - **value of feature for region *R*:** $\hat{p}(w \in R) = \frac{n_w + 1}{n_w + (1/f_R)}$
    - **Bayesian estimator**
    - **beta prior equiv to:**
      - uniform prior + 1 observation in *R* out of $1/f_R$ observation of word
    - **chosen because:**
      - prior = frequency of randomly selected word
      - prior not greatly affected for low freq. words not occurring in *R*
      - pushed strongly towards empirical freq. for words occurring in *R*

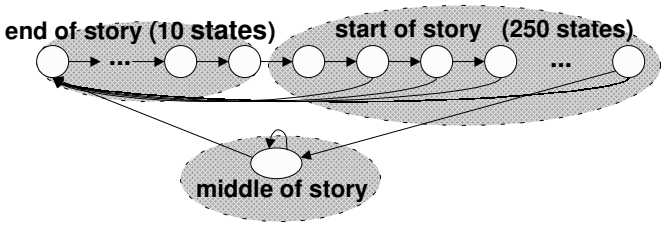
Trigger Features (cont.)

- **Submission run:**
  - . **1st**
  - . **2nd**
  - . **last**
  - . **next-to-last**

  } word of story

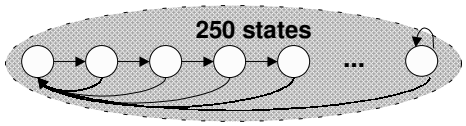
- **initial experimentation with other regions:**
  - **unfruitful, so far**

HMM: reduced end-of-story states



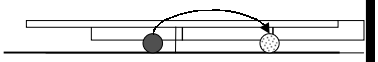
- **Problem with previous model**
  - **confusion: end of short story / end of long story**
  - **reduces value of coherence features**
- **10 end states preserved for: end-of-story trigger cue**

HMM - current version: no end-of-story states

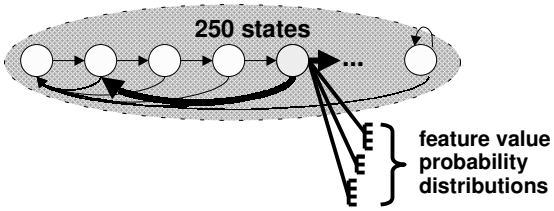


- **Problem with previous model**
  - **confusion still degrades performance**
  - **reduces value of coherence features**
- **end-of-story triggers:**
  - **delay observation for 10 counts:**
    - e.g: {-4,-3,-2} region:
    - high values can be interpreted as:

      "a word showing preference for occuring 2-4 word prior to end of story, ocurred 10 words ago"
    - high values most likely in states 6,7,8

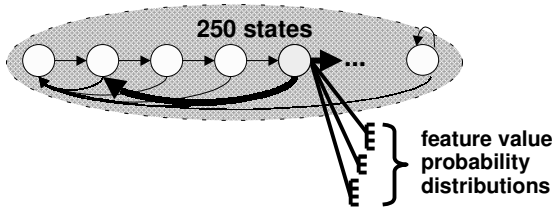


Probability estimation problem:



- **probabilities must be** **estimated from training data**
- **transition probabilities for each state**
- **observation probabilities**
  - **for each feature**
    - **for each possible value of observation**
    - **in each state**

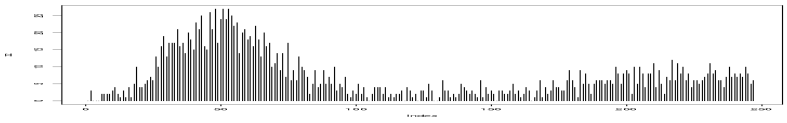
Probability estimation - smoothing:



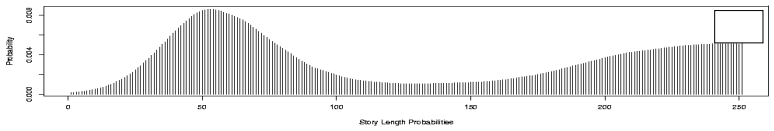
- assumptions:
  - transition probabilities don't vary much
    - from state to state
  - for a given feature, observation probabilities should be similar:
    - for neighboring values
    - in neighboring states

Parameter estimation - transition probabilities

- Raw histogram: ABC_WNT



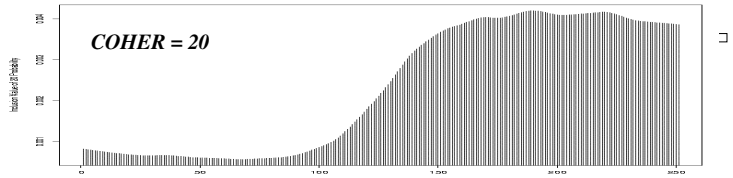
- Smoothed density estimation
  - in place of raw histogram



- from this, cond. probs: $p(S_{t+1} = 0 | S_t = s)$ derived directly

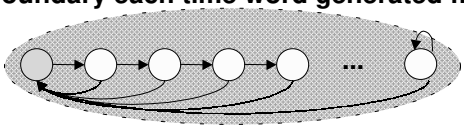
Non-parametric estimation: observation probs

- values are binned
- for each binned feature value, *fv*:
  - non-parametric regression: *fv-count* vs. *state*
  - normalize smoothed counts => *p(fv | s)*
  - *p(fv | s) * p(s) => p(fv, s)*
- for each state, *s*:
  - normalize *p(fv, s)* across values => *p(fv | s)*

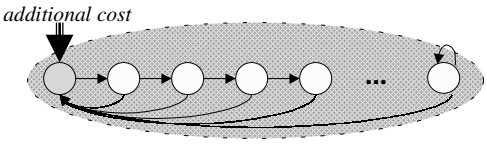


Segmentation

- Viterbi algorithm: Maximum Likelihood state sequence
- output boundary each time word generated from state=1

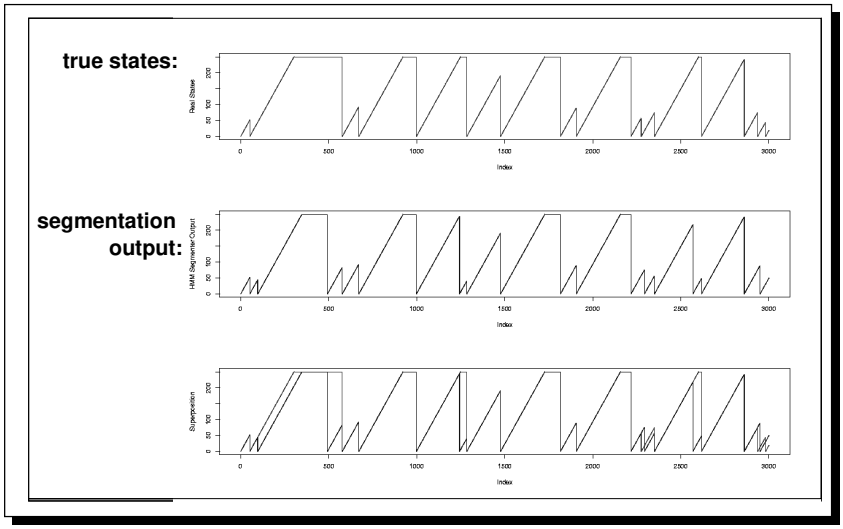


- boundary state "penalty":

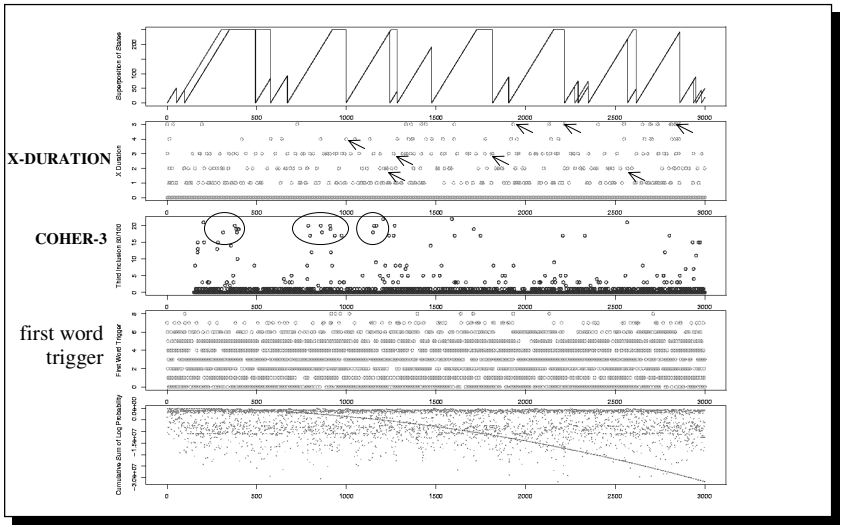


  - tuning for evaluation metric

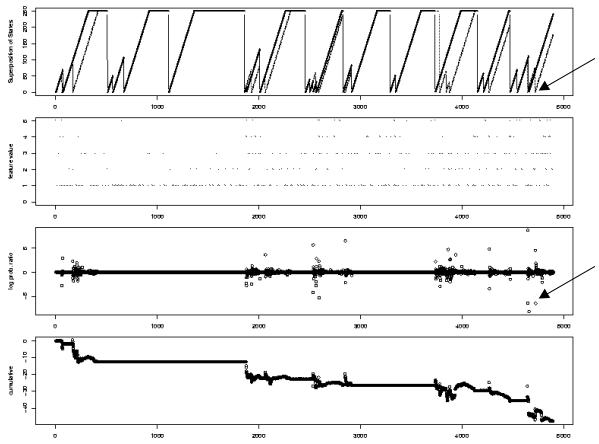
segmentation performance (ABC_WNT)



Visualization: effect of observation values on segmentation



Visualization: effect of X-DURATION



Results

- **ABC_WNT:**
  - **encouraging**
  - **train on all but 15 files**
  - **test of remaining 15**
  - **.31 (TDT-3:** IBM=.31; CMU=.33**)**
- **approach needs to be applied to other sources**

Future work

- **Extension to other sources**
- **Theoretically motivated methods for evaluating**
  - feature potential
  - impact on segmentation performance over multiple files
- **Smoothing**
  - degree of polynomial, kernel function, bandwidth
  - goodness-of-fit
- **Features:**
  - textual
  - audio signal