Chapter 9:

Topic Detection and Tracking

(TDT)

Some slides from “Overview NIST Topic Detection and Tracking
-Introduction and Overview” by G. Doddington
-http://www.itl.nist.gov/iaui/894.01/tests/tdt/tdt99/presentations/index.htm

TDT Task Overview

- 5 R&D Challenges:
  - Story Segmentation
  - Topic Tracking
  - Topic Detection
  - First-Story Detection
  - Link Detection

- TDT3 Corpus Characteristics:†
  - Two Types of Sources:
    - Text
    - Speech
  - Two Languages:
    - English stories 30,000
    - Mandarin stories 10,000
  - 11 Different Sources:
    - _8 English__ 3 Mandarin

8 English		3 Mandarin
ABC	CNN	VOA
PRI	VOA	XIN
NBC	MNB	ZBN
APW	NYT	

* see http://www.itl.nist.gov/iaui/894.01/tdt3/tdt3.htm for details

† see http://morph.ldc.upenn.edu/Projects/TDT3/ for details

Preliminaries

A **topic** is …

a seminal **event** or activity, along with all directly related events and activities.

A **story** is …

a topically cohesive segment of news that includes two or more DECLARATIVE independent clauses about
a single event.

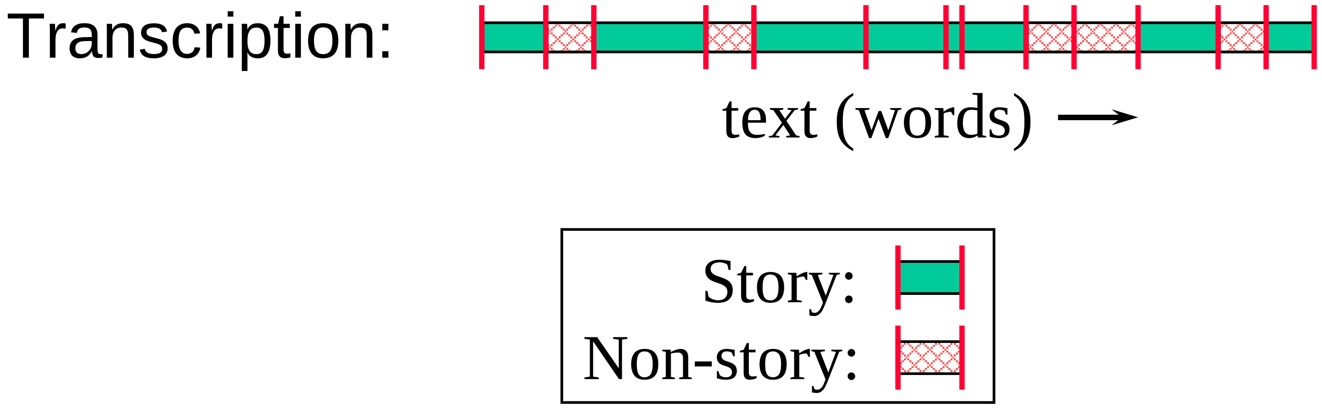
Example Topic

Title: Mountain Hikers Lost

- WHAT: 35 or 40 young Mountain Hikers were lost in an avalanche in France around the 20th of January.
- WHERE: Orres, France
- WHEN: January 1998
- RULES OF INTERPRETATION: 5. Accidents

The Segmentation Task:

To segment the source stream into its constituent stories, for all audio sources.

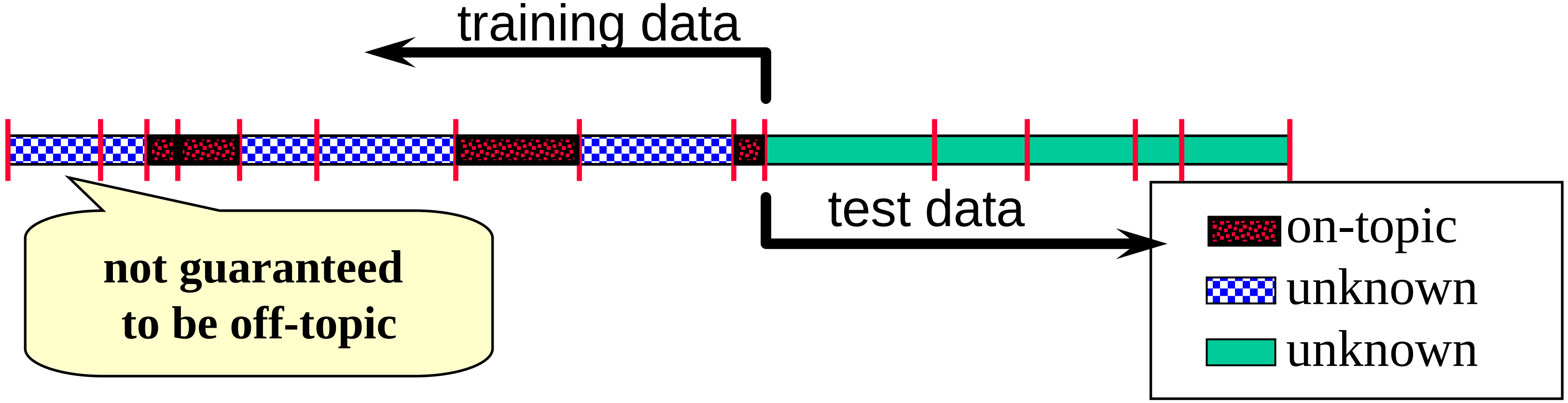


(for Radio and TV only)

The Topic Tracking Task:

To detect stories that discuss the target topic, in multiple source streams.

- Find all the stories that discuss a given target topic
  - *Training:* Given N_t sample stories that discuss a given target topic,
  - *Test:* Find all subsequent stories that discuss the target topic.

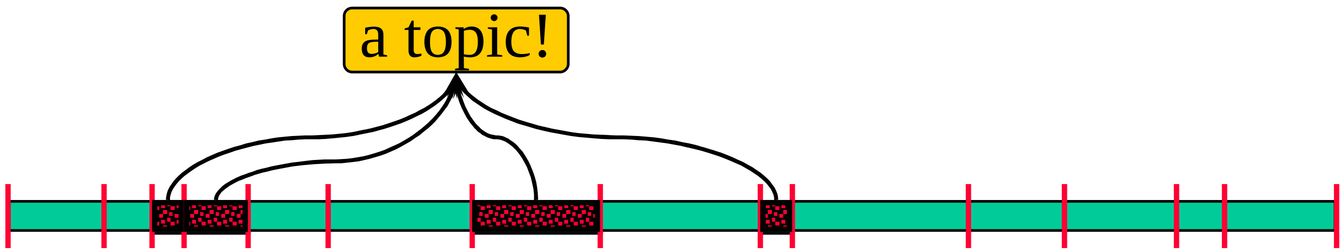


Topic Tracking Conditions

- 3 Source Conditions:
  - text sources and manual transcription of the audio sources
  - text sources and ASR transcription of the audio sources
  - text sources and the sampled data signal for audio sources
- 2 Story Boundary Conditions:
  - Reference story boundaries provided
  - No story boundaries provided

The Topic Detection Task:

To detect topics in terms of the (clusters of) stories that discuss them.



- Unsupervised topic training
- New topics must be detected as the incoming stories are processed.
- Input stories are then associated with one of the topics.

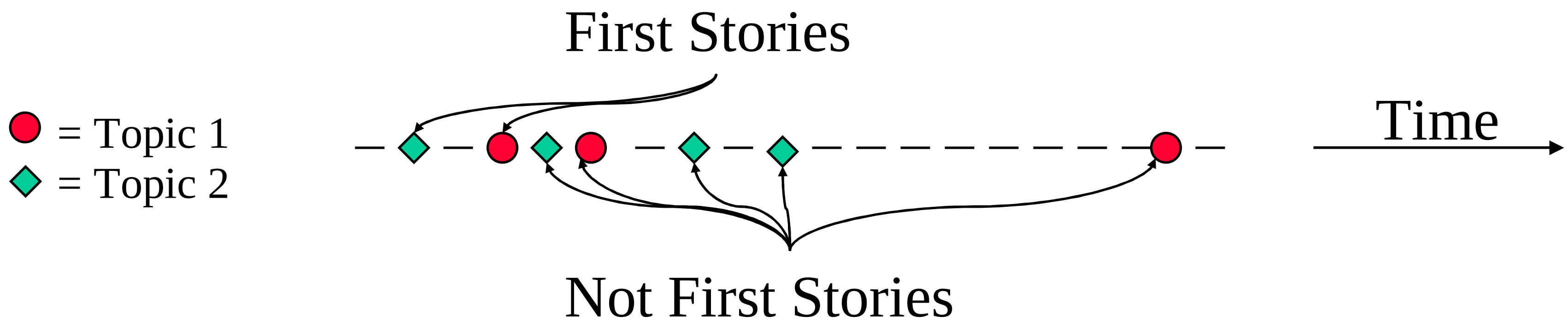
Topic Detection Conditions

- Decision Deferral Conditions:

Maximum decision deferral period in # of source files
1
10
100

The First-Story Detection Task:

To detect the first story that discusses a topic, for all topics.

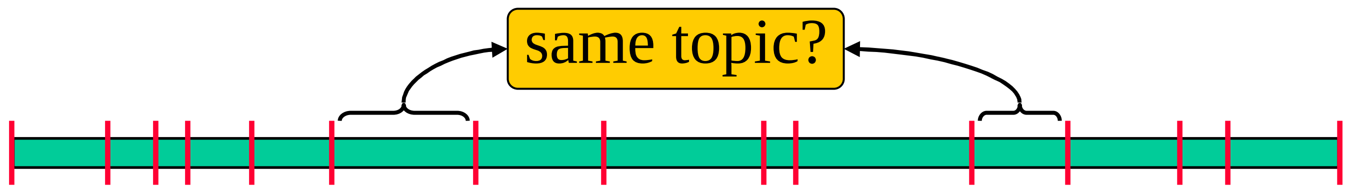


- There is no supervised topic training

  (like Topic Detection)

The Link Detection Task

To detect whether a pair of stories discuss the same topic.



- The topic discussed is a free variable.
- Topic definition and annotation is unnecessary.
- The link detection task represents a basic functionality, needed to support all applications (including the TDT applications of topic detection and tracking).
- The link detection task is related to the topic tracking task, with Nt = 1.

TDT3 Evaluation Methodology

- All TDT3 tasks are cast as statistical detection (***yes*-*no***) tasks.
  - Story Segmentation: ***Is there a story boundary here?***
  - Topic Tracking: ***Is this story on the given topic?***
  - Topic Detection: ***Is this story in the correct topic-clustered set?***
  - First-story Detection: ***Is this the first story on a topic?***
  - Link Detection: ***Do these two stories discuss the same topic?***
- Performance is measured in terms of detection cost, which is a weighted sum of ***miss*** and ***false alarm*** probabilities:

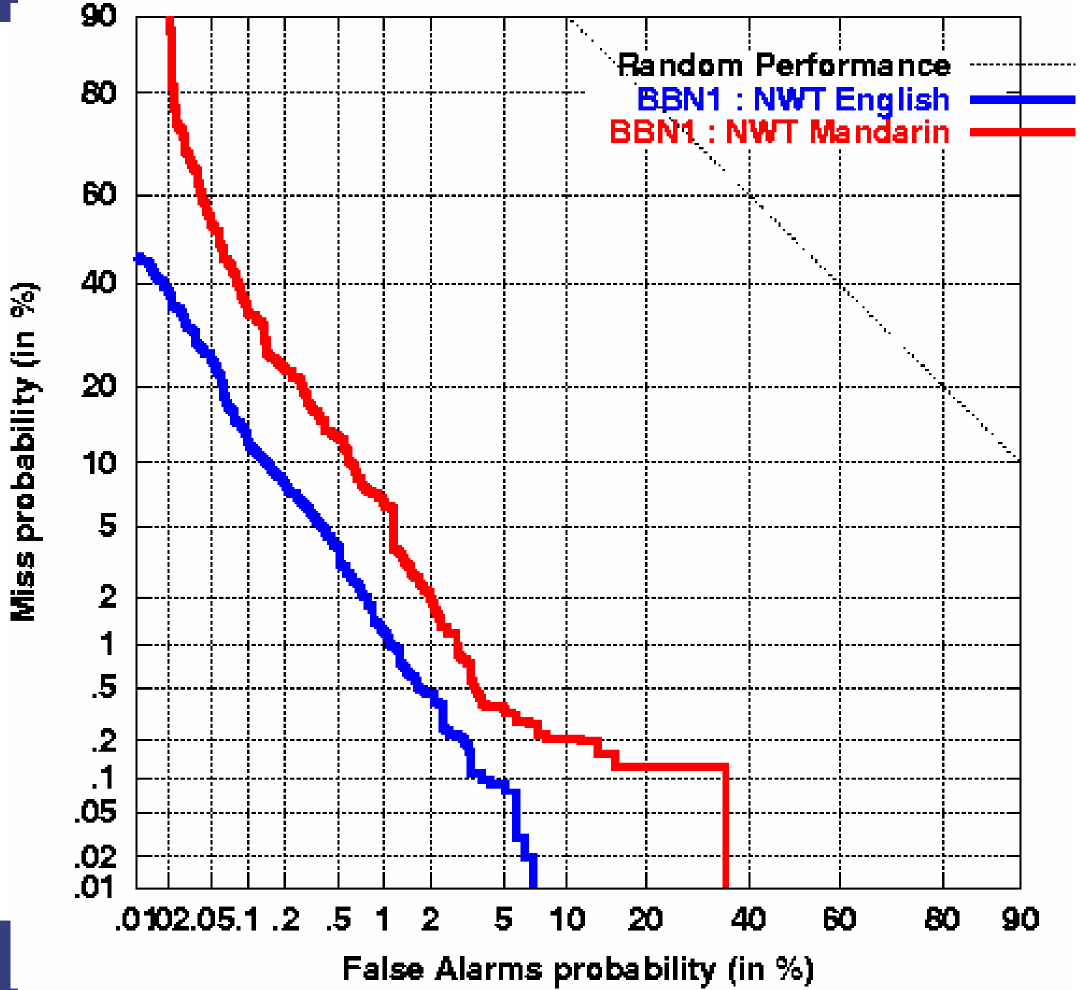
  $$\mathbf{C_{Det}} = C_{Miss} \bullet P_{Miss} + C_{FA} \bullet P_{FA}$$

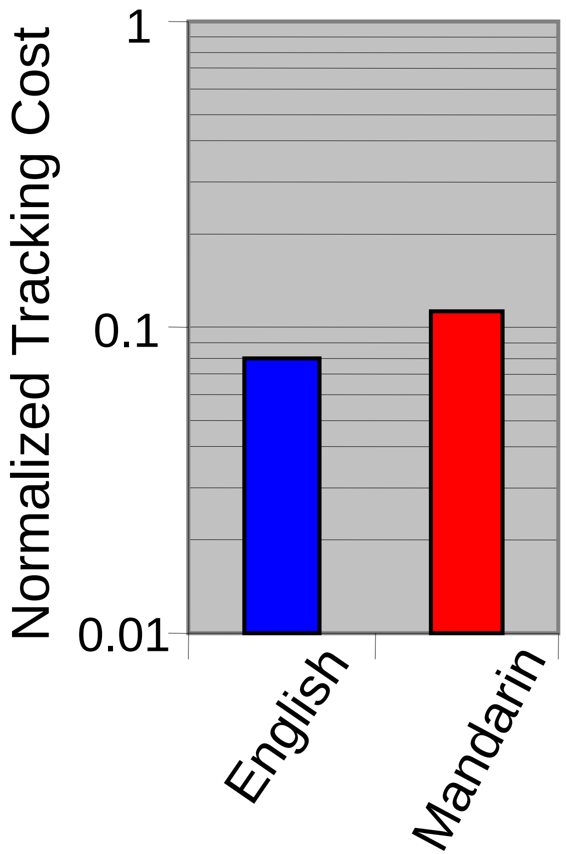
  (e.g. $C_{Miss}=0.2$, $C_{FA} = 0.98$)
- Detection Cost is normalized to lie between 0 and 1:

  $$\mathbf{(C_{Det})_{Norm}} = \mathbf{C_{Det}} \,/\, \min\{C_{Miss},\ C_{FA}\}$$

Example Performance Measures:

Tracking Results on Newswire Text (BBN)

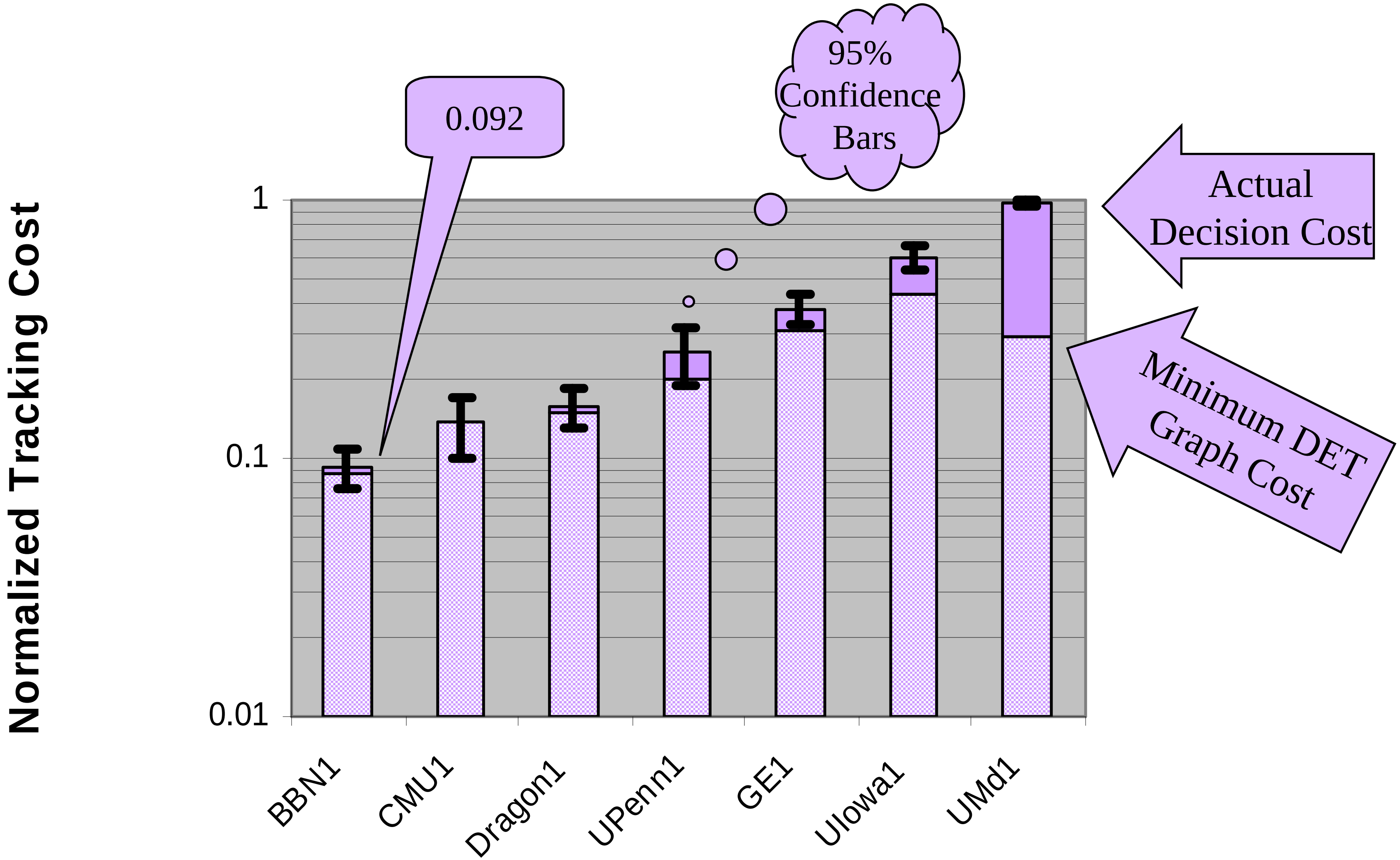




1999 TDT3 Tracking Results

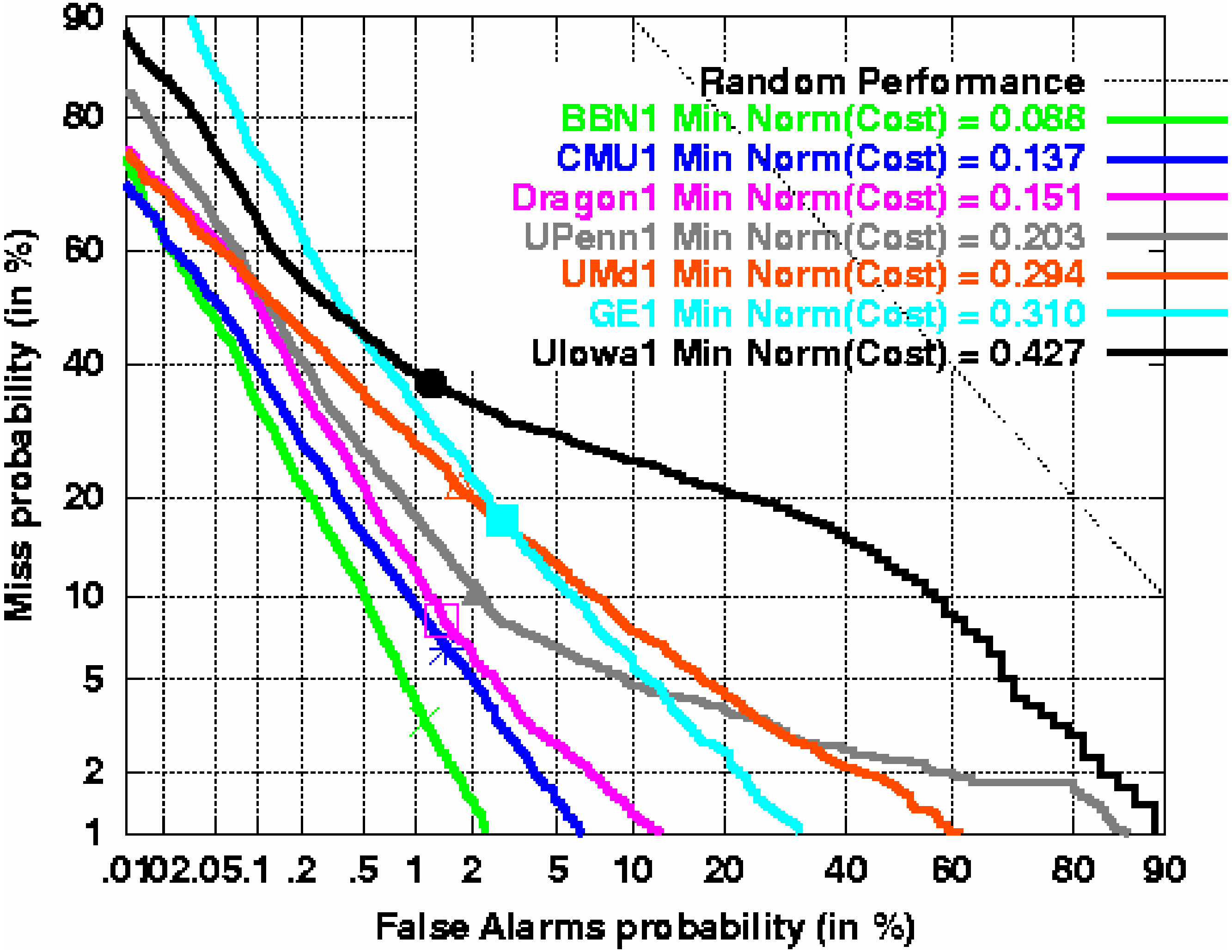
Required Evaluation Condition

4 English Training Stories, Multilingual Test Texts,
Newswire Text+Broadcast News ASR, Given ASR Boundaries

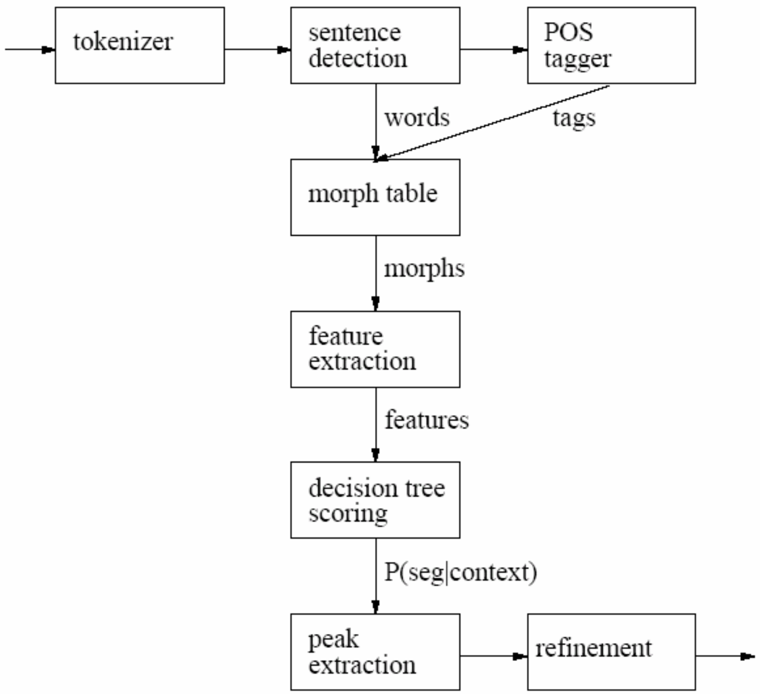


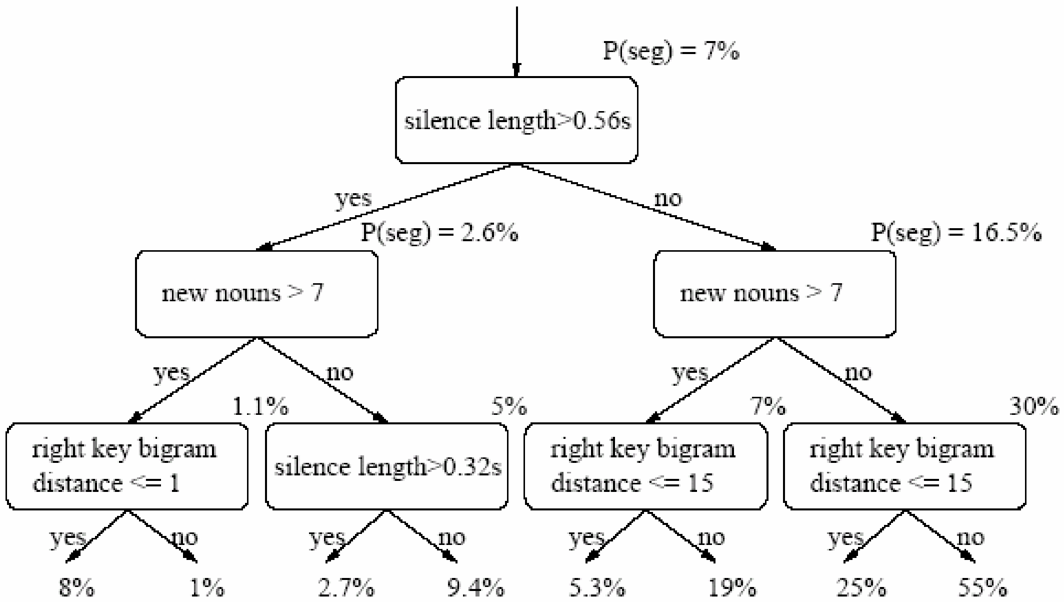
1999 TDT3 Tracking Results

Required Evaluation Condition



Story Segmentation using Decision Trees





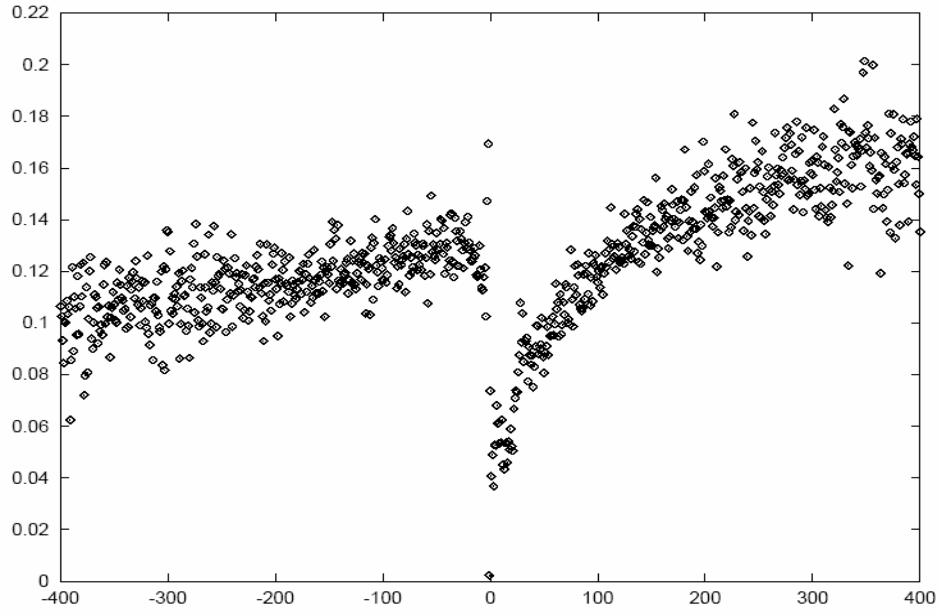
Story Segmentation and Topic Detection in the Broadcast News Domain

S. Dharanipragada *M. Franz* *J.S. McCarley* *S. Roukos* *T. Ward*

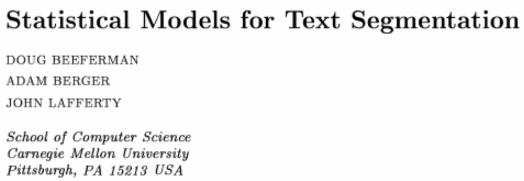
IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Using Maximum Entropy Language Models

Idea: compare perplexities of adaptive trigram with general English trigram



Relative position in segment



Results

segmentation model	P_k	*miss probability*	*false alarm probability*
exponential model	13.2%	16.0%	10.9%
decision tree	15.2%	19.3%	11.9%
interpolated (exp + dtree) models	11.8%	14.2%	9.8%
cue-word and $s = t$ trigger features	13.4%	16.9%	10.5%
cue-word and $s \neq t$ trigger features	13.6%	17.8%	10.1%
cue-word features only	18.3%	21.6%	15.5%
topicality features only	37.3%	42.1%	33.3%
TextTiling	34.6%	57.1%	18.6%

Relevance Models and Link Detection

- Given two stories A and B
  - Determine if topic(A)=topic(B)
- Estimate topics models of A and B
  - e.g. language models
- Measure distance between the models
  - e.g. Kullback-Leibler

Relevance Models for Topic Detection and Tracking

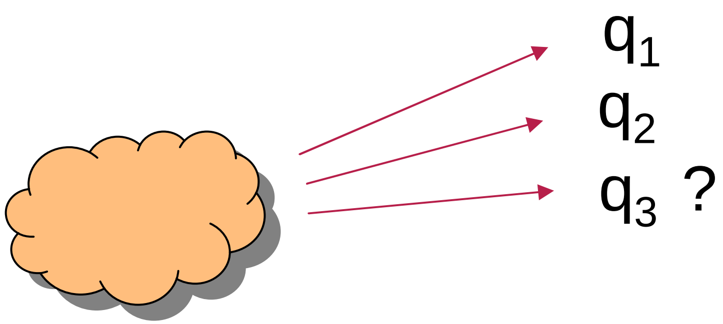
Victor Lavrenko, James Allan,
Edward DeGuzman, Daniel LaFlamme, Veera Pollard, and Stephen Thomas

Center for Intelligent Information Retrieval
Department of Comupter Science
University of Massachusetts, Amherst, MA 01003

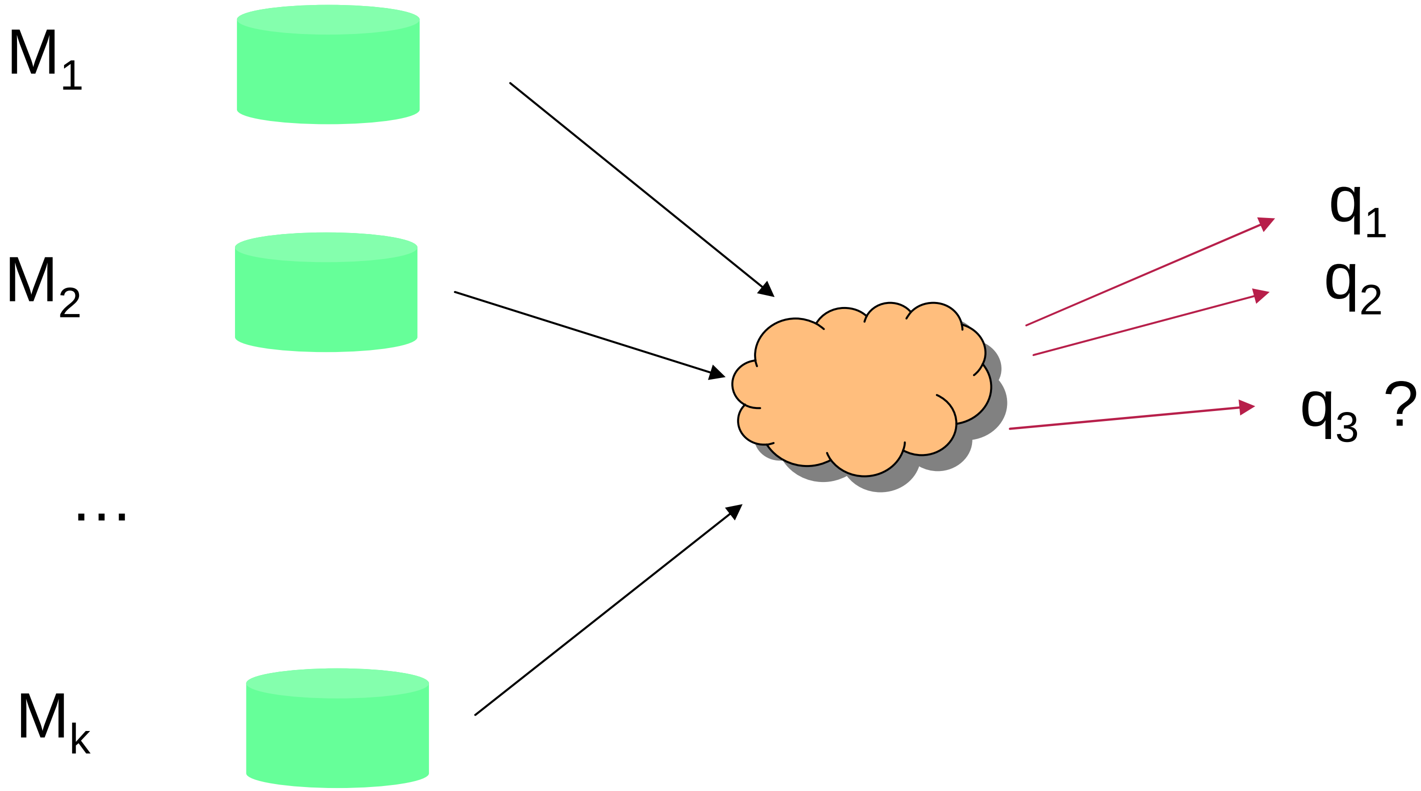
Generating Queries

- Suppose you have some source of queries
- You have generated several queries $q_1 \ldots q_n$ from this source
- What is

$$P(q_{N+1} \mid q_1 \ldots q_{N-1})$$



Universe of Models



$$P(q_{N+1} \mid q_1 \ldots q_N) = \sum_{i=1}^{k} P(q_{N+1} \mid M_i) P(M_i \mid q_1 \ldots q_N)$$

Using Relevance Models in Link Detection

- Question:
  - Are stories S_1 and S_2 linked?
- Approach
  - Create are relevance model for S_1 and S_2
  - Measure the distance between the models

Building Relevance Models as Topic Models

- S_1 :
  - Generate queries from it
  - Retrieve documents from the collection
  - Estimate

$$P(w \mid D) = \lambda \frac{tf_{w,D}}{|D|} + (1-\lambda)\frac{cf_w}{Coll.Size}$$

$$P(w \mid S_1) = P(w \mid q_1 \ldots q_N)$$

$$= \sum_{D \in R} P(w \mid D) P(D \mid q_1 \ldots q_N)$$

Measuring Distances

Kullback-Leibler Distance

$$D(S_1 \| S_2) = \sum_w P(w|S_1) \log \frac{P(w|S_1)}{P(w|S_2)}$$

Symmetric Kullback-Leibler Distance

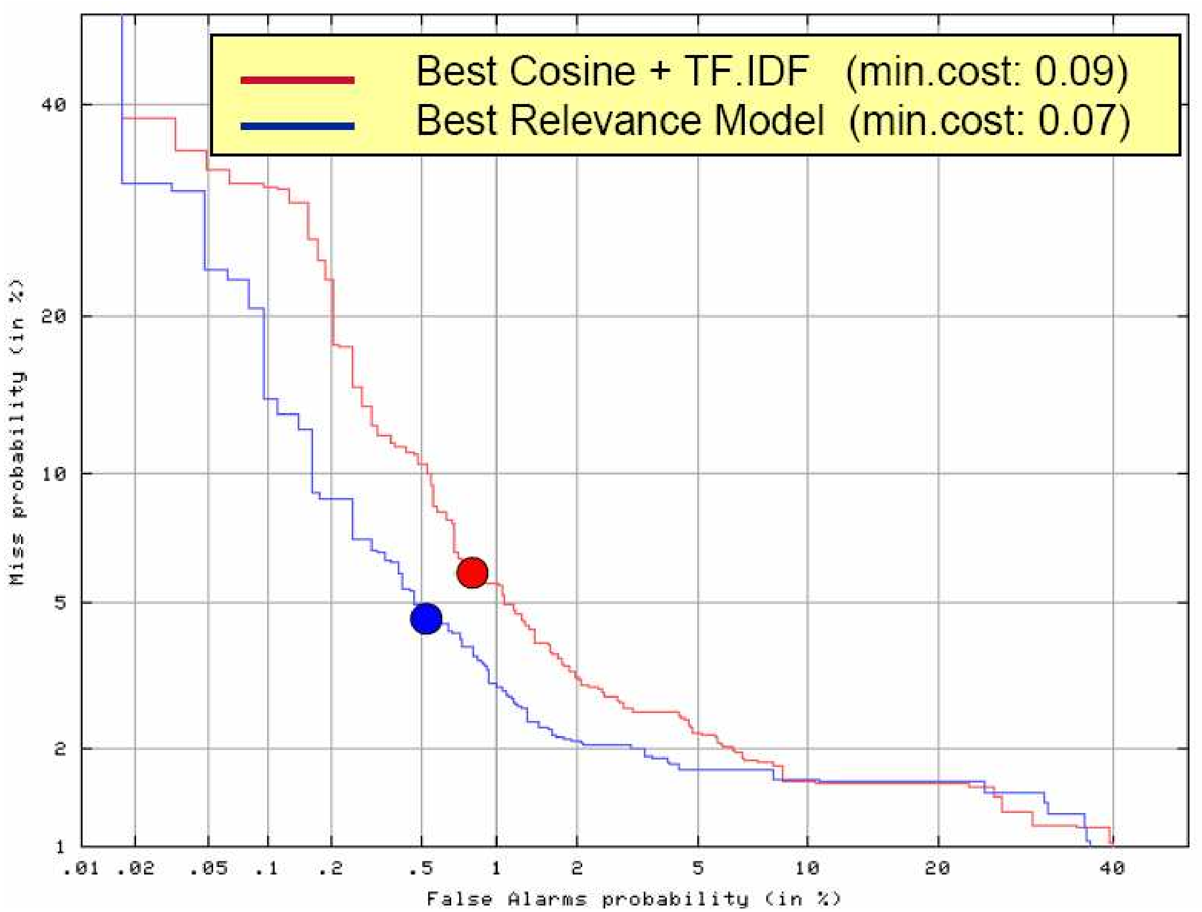
$$D_{sym}(S_1 \| S_2) = \frac{1}{2}\left(D(S_1 \| S_2) + D(S_2 \| S_1)\right)$$

Kullback-Leibler Distance with "Clarity"

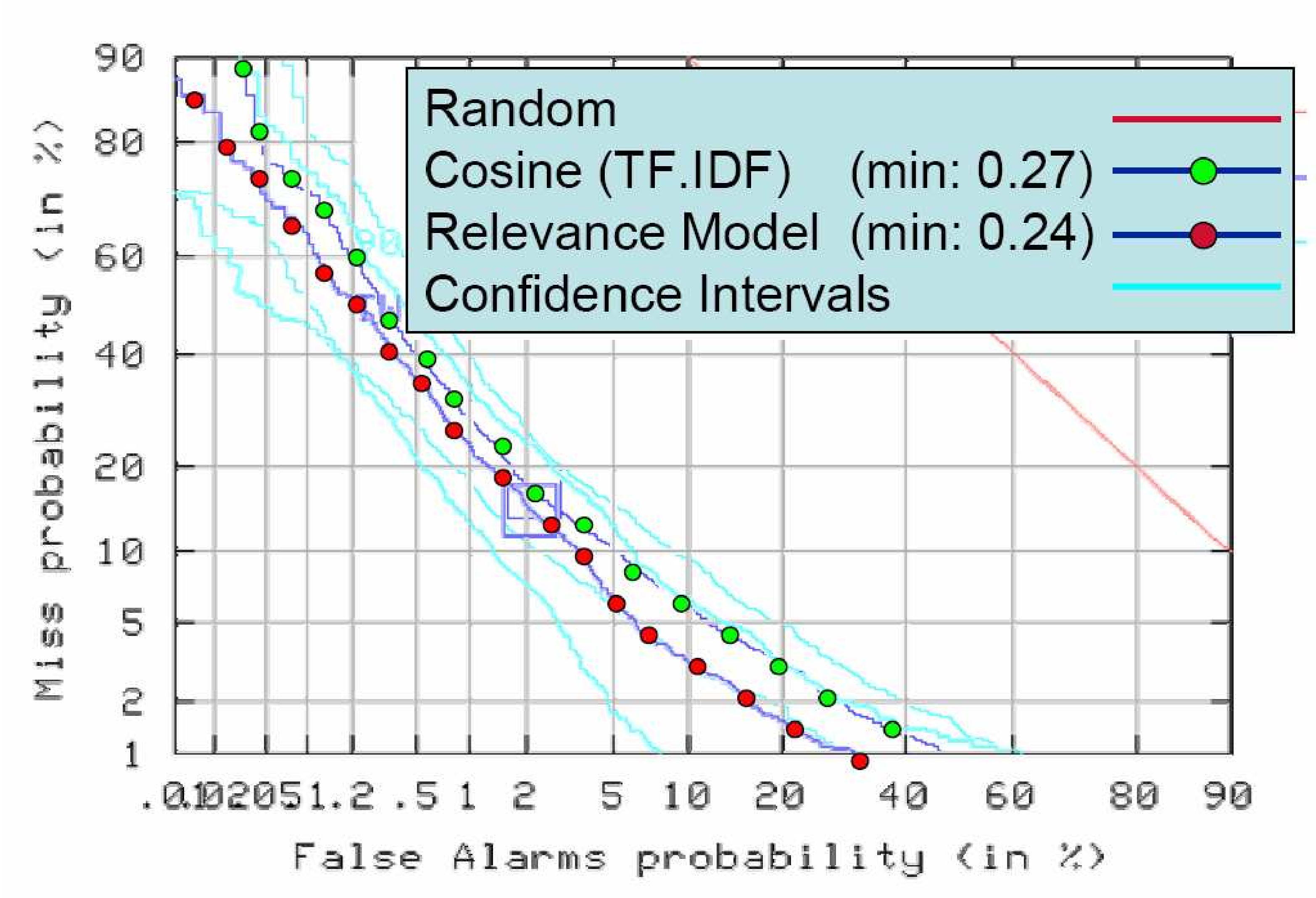
$$D_{Cl}(S_1 \| S_2) = \sum_w P(w|S_1) \log \frac{P(w|S_2)}{P(w|GE)}$$

$$(\mathrm{GE} : \mathrm{general\ english})$$

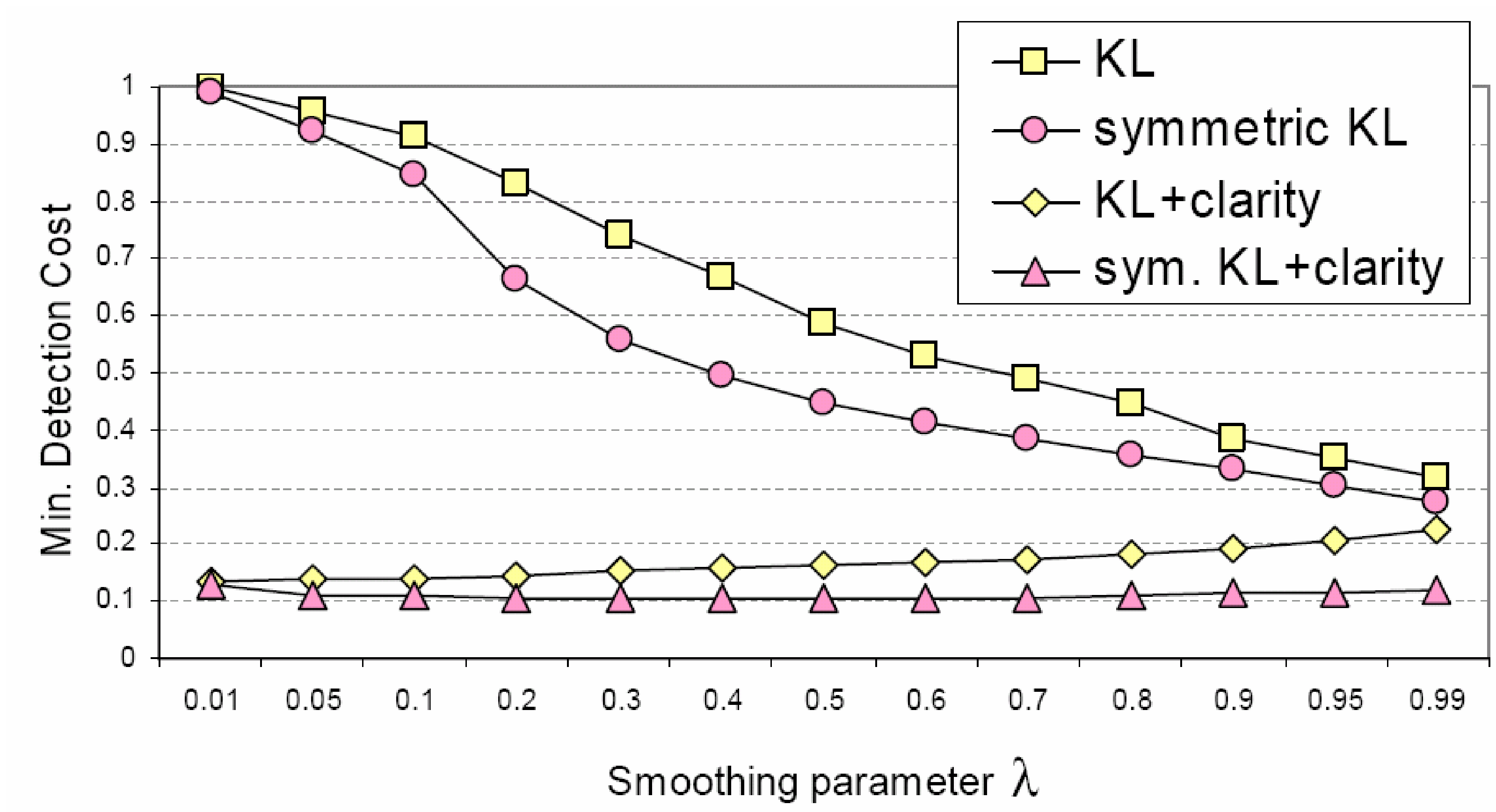
Comparison on Training Data



Comparison on Evaluation Data



Distance Metric /Smoothing



Sym. KL distance + clarity is not only the best method but also is robust against changes in the smoothing

Summary

- TDT:
  - International Benchmark
  - Various sub tasks
- Link detection