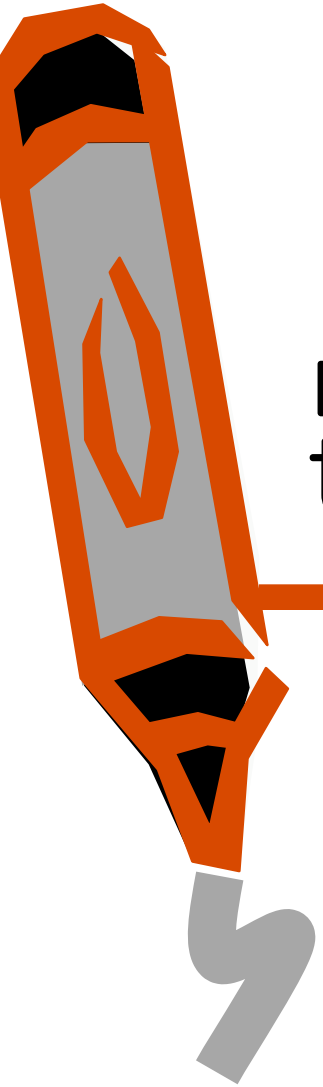


news filtering
topic detection and tracking

some slides (c) James Allan@umass
some slides (c) Ray Larson @University of California, Berkeley
some slides (c) Jian Zhang and Yiming Yang@Carnegie Mellon University
some slides (c) Christopher Cieri @University of Pennsylvania

outline

- news filtering
- TDT
- advanced TDT
- novelty detection

Google news

Sign in

Google News

Web | Images | Groups | **News** | Froogle | Maps | more » | Advanced News Search

Search News | Search the Web

Search and browse 4,500 news sources updated continuously.

Standard News | Text Version

>Top Stories
World
U.S.
Business
Sci/Tech
Sports
Entertainment
Health
Most Popular

News Alerts

RSS | Atom
About Feeds

Mobile News

Top Stories U.S. Go

Auto-generated 3 Aug at 4:24 GMT

Castro's Younger Brother Is Focus of Attention Now

New York Times - **2 hours ago**

With the mysterious illness of Fidel Castro this week, attention has turned to his brother, Raúl, the new provisional leader of Cuba, a man whose personality is little known to Cubans and who remained out of sight and silent on Wednesday. ...

Castro seeks to reassure Cubans of good health Canada.com

Hope is for 'beginning of the end' Palm Beach Daily News

Christian Science Monitor - TheOnlinewire - Voice of America - abc7.com - **all 3,916 related »**

Central Florida News 13

Tropical Storm Chris Prompts Hurricane Watch for Bahamas

Voice of America - **9 hours ago**

By VOA News. The Bahamas has issued a hurricane watch Wednesday for the Turks and Caicos islands and southeastern Bahamas, as Tropical Storm Chris continues its path through the Caribbean. US forecasters say ...

5 pm update - Chris weakens slightly Sun-Sentinel.com

Cuba or the Keys for Chris? Houston Chronicle

SooToday.com - Charlotte Sun-Herald - Caribbean360.com - Shreveport Times - **all 1,303 related »**

Caribbean360.com

Get recommended stories

Sign in to get recommended stories by using search history

Personalize this page

AOL Casts Its Fate With Ads
BusinessWeek - all 863 related »

IBM joins the AMD Opteron choir
VNUNet.com - all 179 related »

Penny pushes Dodgers to fifth straight
MLB.com - all 239 related »

Gibson Is Charged With Misdemeanors
Los Angeles Times - all 1,786 related »

Despite Action on Plan B, FDA Nominee Is in Limbo
New York Times - all 994 related »

In The News

Fidel Castro
Time Warner
Floyd Landis
Tony Blair
Eastman Kodak
Cadbury Schweppes
West Nile
Steve McClaren
Devil Rays
American Legion

World »

edit

Olmert Denies Reports Of Temporary Cease-Fire

Evening Bulletin - **13 hours ago**

Jerusalem - Prime Minister Ehud Olmert last night denied reports of a temporary cease-fire with Hezbollah, firmly asserting that the fighting would continue. Olmert spoke at a meeting of local authority leaders ...

200 missiles hit Israel as battle rages in Lebanon International Herald Tribune

Heavy equipment used to bury the dead Houston Chronicle

RTE.ie - New York Times - Chicago Tribune - Bloomberg - **all 1,942 related »**

Global National

Activist party leader wants fresh Rada elections

Kyiv Post - **4 hours ago**

by Lena Plekhanova, Kyiv Post Staff Writer. PORA, a youth organization that gained fame during the Orange Revolution by providing security for pro-democracy street demonstrations, before running as a political

U.S. »

edit

Minimum wage-estate tax bill snagged in US Senate

Reuters - **6 hours ago**

By Richard Cowan. WASHINGTON, Aug 2 (Reuters) - The fate of an election-year bill to raise the federal minimum wage for low-income Americans and cut inheritance taxes for the wealthiest is in doubt in the Senate ...

Durbin leads fight against tax-cut bill Peoria Journal Star

Candidates spar over minimum wage bill Newsday

KWWL - Forbes - Investor's Business Daily (subscription) - SitNews - **all 605 related »**

Washington Post

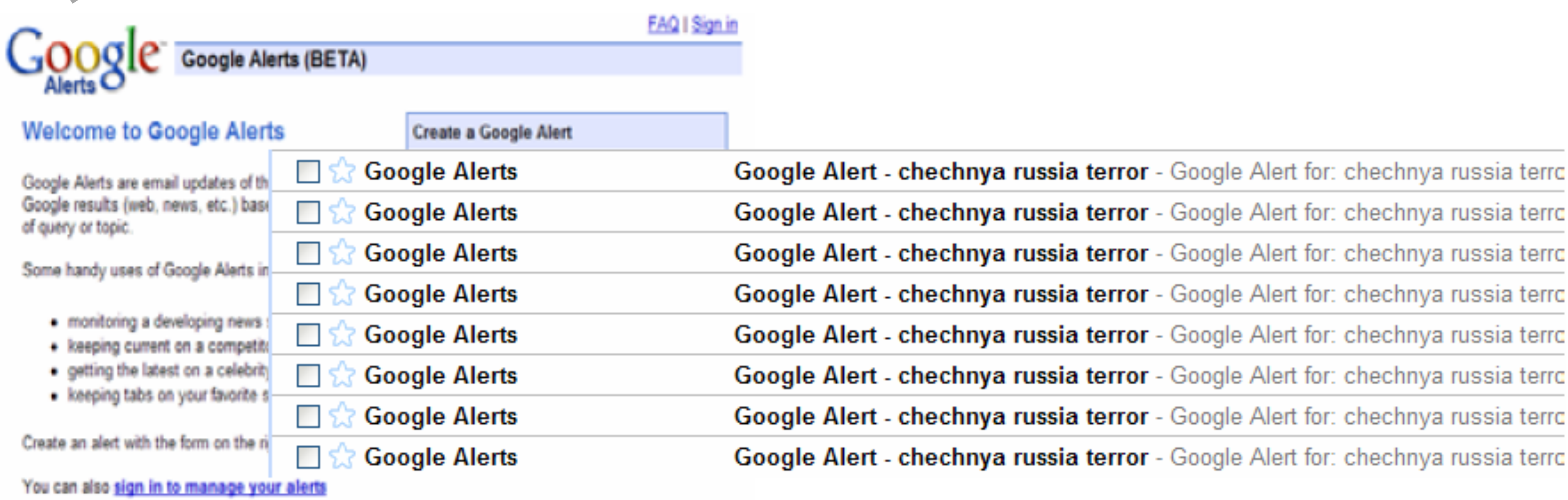
Marine generals: Haditha probe helps reinforce corps' integrity

Myrtle Beach Sun News - **7 hours ago**

CAMP LEJEUNE, NC - The departing and incoming commanders of the 2nd Marine Expeditionary Force believe the investigation into accusations that US Marines deliberately shot civilians in Haditha, Iraq, will ultimately strengthen the corps

CNN

Google alerts



RSS feeds

http://www.washingtonpost.com/wp-dyn/rss/index.html#what

- XML feeds
- Lots of News sites provide it now
- Web content providers can easily create and disseminate feeds of data that include news links, headlines, and summaries

news filtering

- TDT and TREC.
- Usually the starting point is a few example documents on each topic.
- TDT topics are events in news.
- TREC topics are broader.
- TREC gives room for user feedback. New feature in TDT.
- Some of the assumptions are unrealistic.

TDT

- Intended to automatically identify new topics – events, etc.
  - from a stream of text and follow the development/ further discussion of those topics
- Automatic organization of news by events
  - Wire services and broadcast news
  - Organization on the fly--as news arrives
  - No knowledge of events that have not happened
- Topics are event-based topics
  - Unlike subject-based topics in IR (TREC)

TDT Task Overview

TDT Task Overview

- 5 R&D Challenges:

TDT Task Overview

- 5 R&D Challenges:
  - Story Segmentation

TDT Task Overview

- 5 R&D Challenges:
  - Story Segmentation
  - Topic Tracking

TDT Task Overview

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

TDT Task Overview

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

TDT Task Overview

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

TDT

- Topic Detection and Tracking
  - focused on detecting and tracking events in news
    - novelty detection: does this story discuss a new event
    - topic tracking: given an example story, track it through time
    - topic detection: organize news stories as they come in
  - targets automatically-recognized radio, TV broadcasts

- Different evaluation: Misses and False Alarms
- Impact of recognition errors:
  - Topic Tracking: minimal
  - Novelty Detection: quite sensitive (unusual problem)
  - very sensitive to absence of story boundaries!

TDT3 corpus

TDT3 corpus

- TDT3 Corpus Characteristics:

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech
  - Two Languages:

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech
  - Two Languages:
    - English 30,000 stories

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech
  - Two Languages:
    - English 30,000 stories
    - Mandarin 10,000 stories

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech
  - Two Languages:
    - English 30,000 stories
    - Mandarin 10,000 stories

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech
  - Two Languages:
    - English 30,000 stories
    - Mandarin 10,000 stories

TDT3 corpus

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

TDT3 corpus

- TDT3 Corpus Characteristics:
  - Two Types of Sources:
    - Text
    - Speech
  - Two Languages:
    - English 30,000 stories
    - Mandarin 10,000 stories
  - 11 Different Sources:
    - 8 English: ABC CNN, PRI VOA, NBC MNB, APW NYT
    - 3 Mandarin: VOA, XIN, ZBN

news, TDT

A topic is ...

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

A story is ...

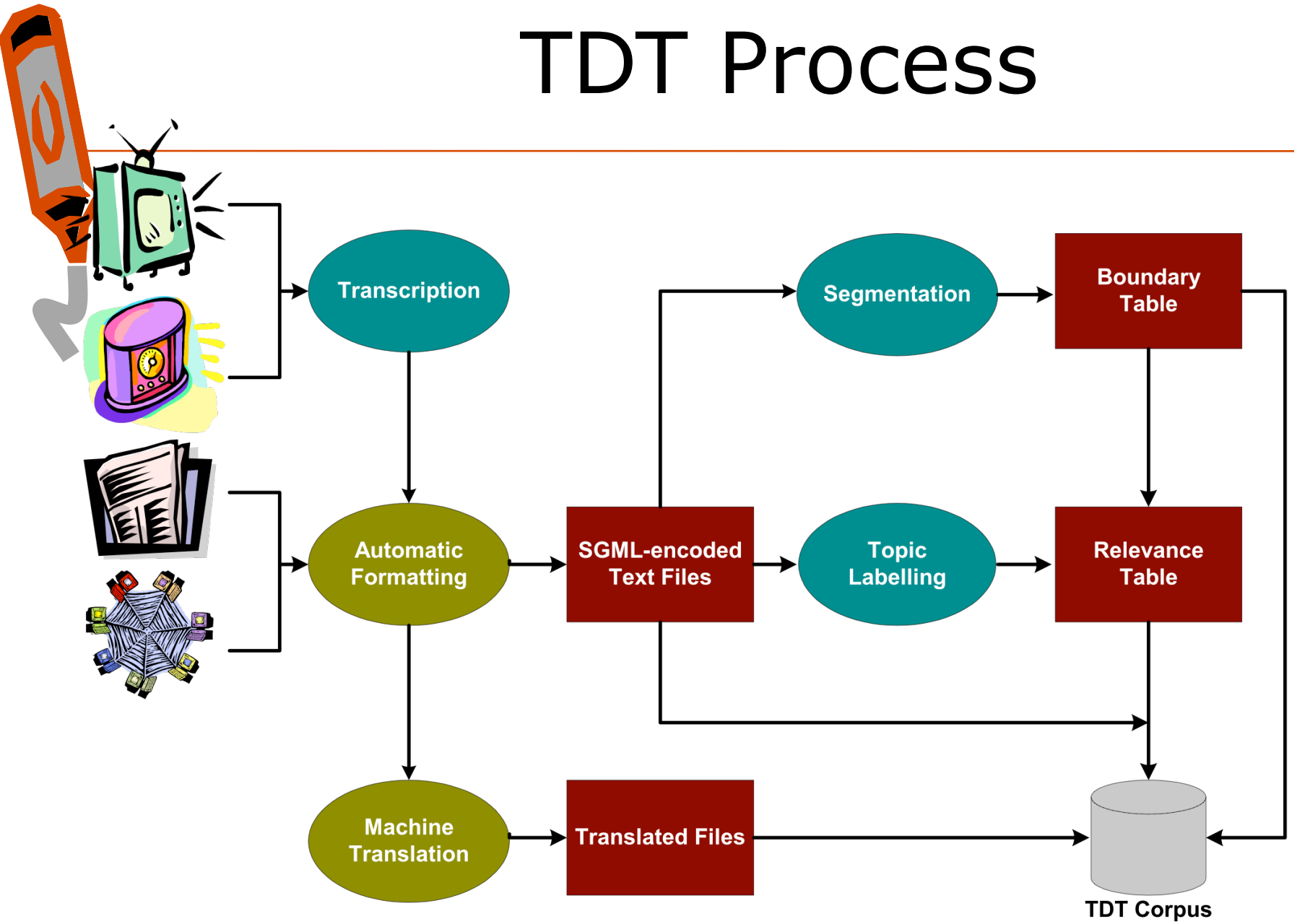
a topically cohesive segment of news that includes two or more DECLARATIVE independent clauses about a single event.

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

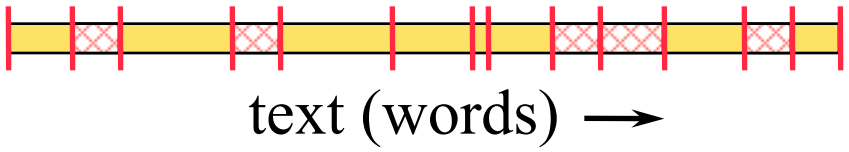
TDT Process

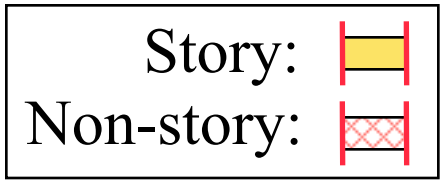

Transcription
Segmentation
Boundary Table
Automatic Formatting
SGML-encoded Text Files
Topic Labelling
Relevance Table
Machine Translation
Translated Files
TDT Corpus

The Segmentation Task:

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

Transcription:





(for Radio and TV only)

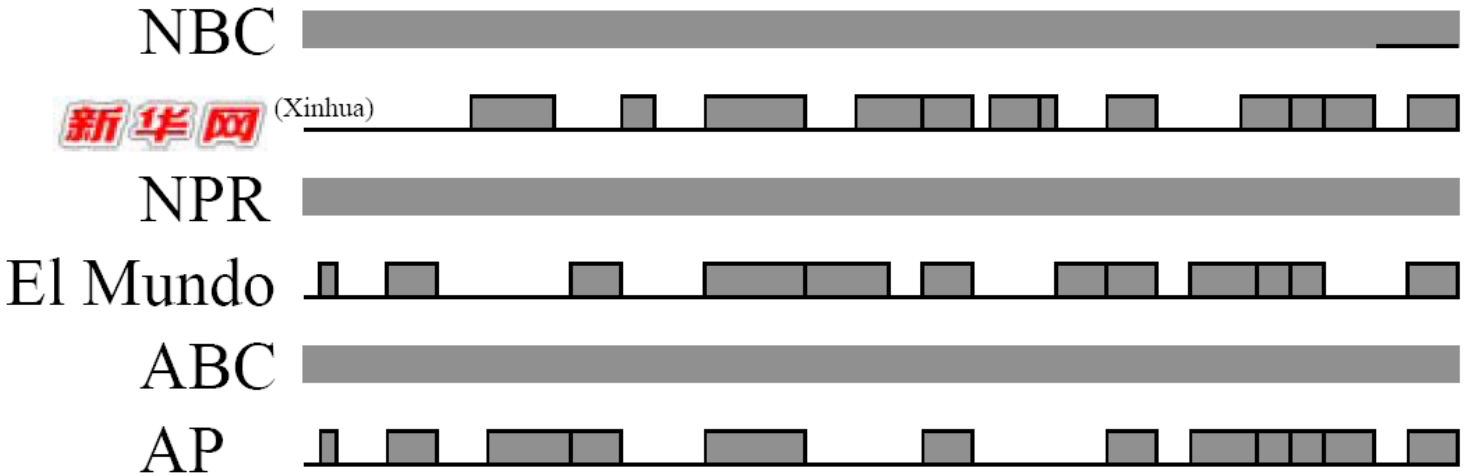
Story Segmentation Conditions

- 1 Language Condition: **Both English and Mandarin**
- 3 Audio Source Conditions:
  - manual transcription
  - **ASR transcription**
  - original audio data
- 3 Decision Deferral Conditions:

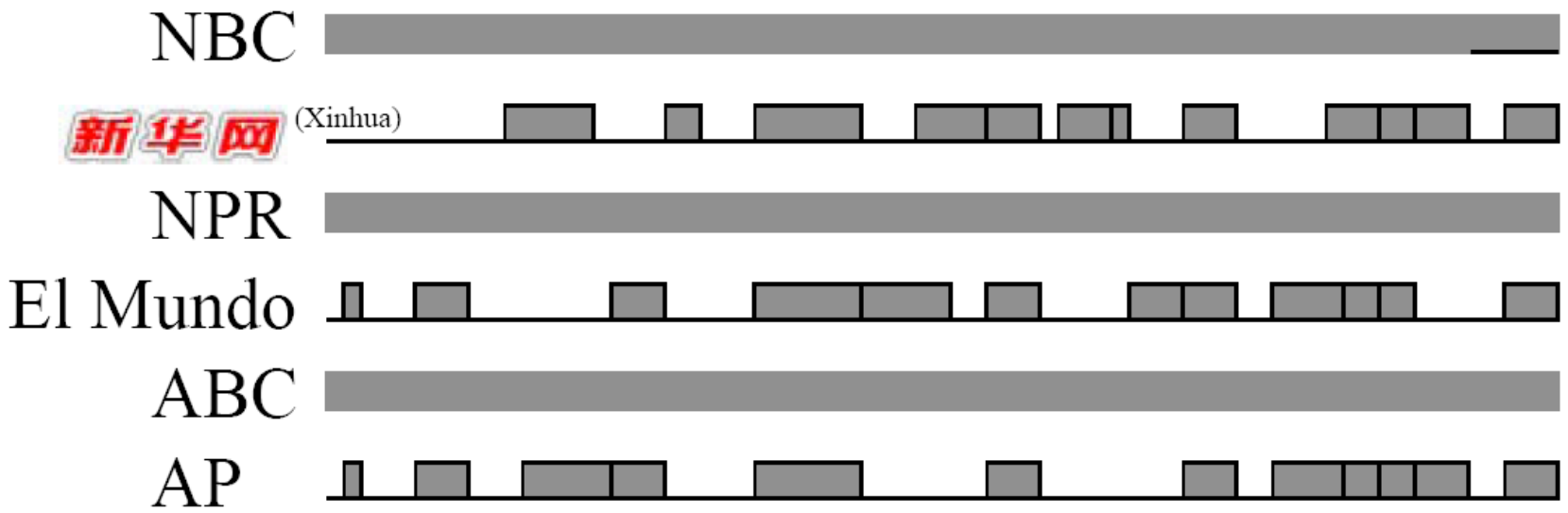
Maximum Decision Deferral Period		
Text		Audio
English (words)	Mandarin (characters)	English & Mandarin (seconds)
100	150	30
1,000	1,500	300
10,000	**15,000**	**3,000**

in reality

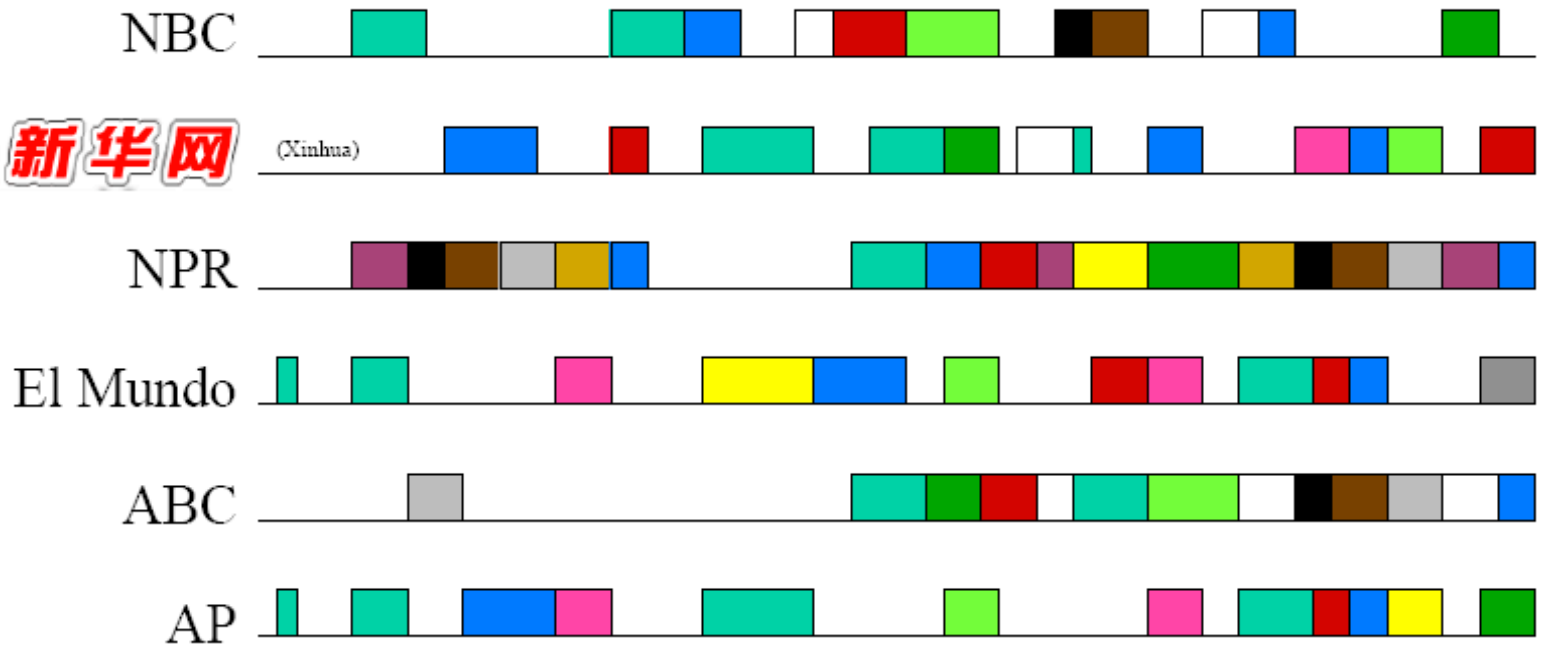
- Events/topics are not given
- Do not know story boundaries for broadcast sources
- Do not know where all of the news is in broadcast sources



TDT : from this



to this



TDT data

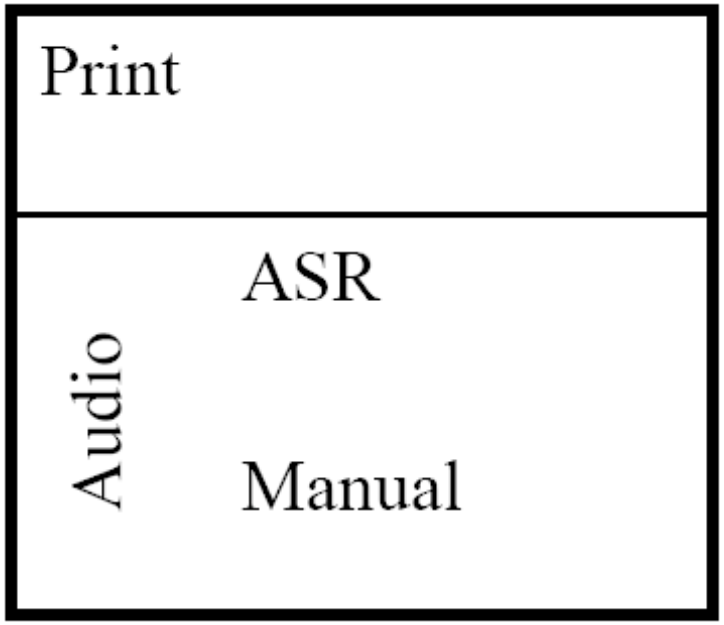
- TDT4 corpus
- Oct 2000 – Jan 2001
- News in Different Languages

English		
Foreign	Mandarin	MT
		Nat
	Arabic	MT
		Nat

Machine Translated
SYSTRAN

TDT data

- TDT4 corpus
- – Oct 2000 – Jan 2001
- News from different sources



TDT data

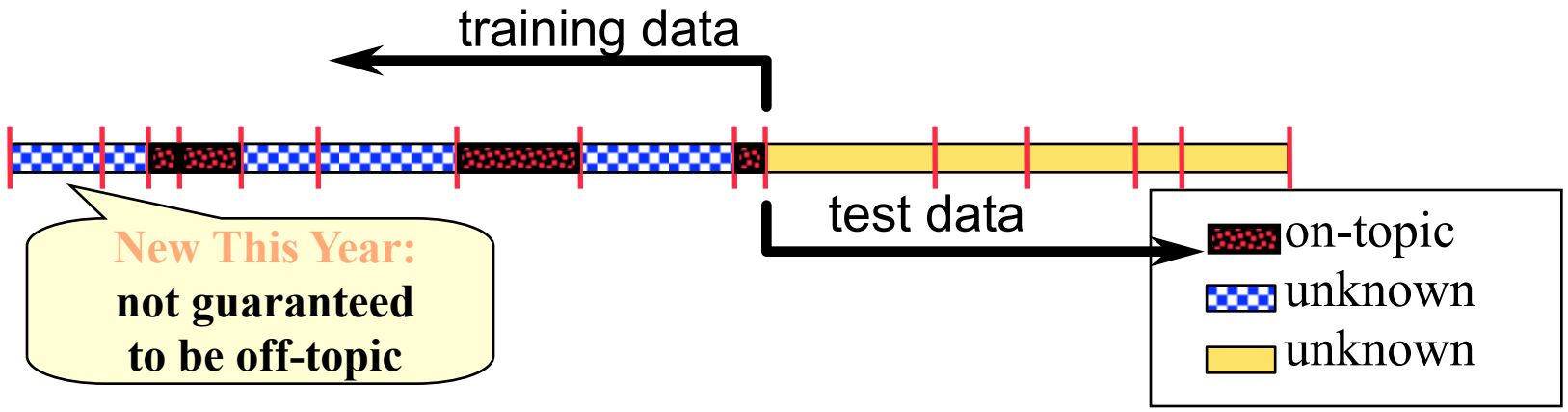
- TDT4 corpus
- – Oct 2000 – Jan 2001
- News from different sources

Print		English	
Audio	ASR	Foreign: Mandarin	MT
			Nat
	Manual	Foreign: Arabic	MT
			Nat

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

- 9 Training Conditions:

Training Language	English	Mandarin	Both Sources
N_t	1 (E)	1 (M)	1 (E), 1(M)
English (E)	2 (E)	2 (M)	2 (E), 2(M)
Mandarin (M)	**4 (E)**	4 (M)	4 (E), 4(M)

- 1 Language Test Condition:

Both English and Mandarin

- 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

topic tracking

<DOC>
<DOCNO> CNN19981002.1600.0051 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 10/02/1998 16:00:51.26 </DATE_TIME>
<BODY>
<TEXT>
new details are out about president clinton's relationship with monica lewinsky. the house judiciary committee has released the last major batch of evidence collected by ken starr in his investigation. the 4,600 pages made public today include transcripts of linda tripp's secret tape recordings of her conversations with lewinsky. testimony by most of the major witnesses who appeared before the grand jury is also included. while this new material doesn't contain the controversial details of previously released documents, it does add color to the contacts between while this new material doesn't contain the controversial details of previously released documents, it does add color to the contacts between tripp and lewinsky.

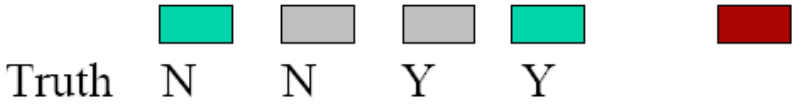
<DOC>
<DOCNO> CNN19981002.1600.0051 </DOCNO>
<DOCTYPE> NEWS </DOCTYPE>
<TXTTYPE> ASRTEXT </TXTTYPE>
<TEXT>
YOU'RE DETAILS ABOUT PRESIDENT CLINTON'S RELATIONSHIP WITH MONICA LEWINSKI TODAY THE HOUSE JUDICIARY COMMITTEE HAS RELEASED THE LAST MAJOR BATCH OF EVIDENCE COLLECTED BY KEN STARR IN HIS SEVEN MONTH PROBE FORTY SIX HUNDRED PAGES MADE PUBLIC TODAY INCLUDE TRANSCRIPTS OF LINDA TRIP SECRET TAPE RECORDINGS OF CONVERSATIONS WITH HER TESTIMONY BY MOST OF THE MAJOR WITNESSES TO APPEAR BEFORE A GRAND JURY IS ALSO INCLUDED WHILE THIS NEW MATERIAL DOESN'T CONTAIN THE CONTROVERSIAL DETAILS OF PREVIOUSLY RELEASED DOCUMENTS IT DOES ADD COLOR THE CONTACTS BETWEEN PRINT AND LOWENSTEIN

the tracking task

- The system is given one training document T_j per story.
- Stories come in sequence $S_1 \ldots S_n$

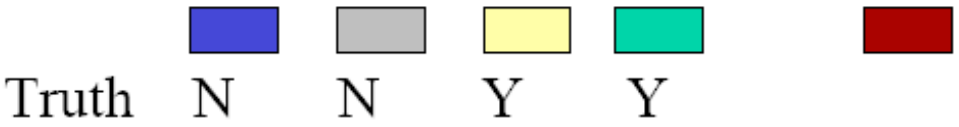
the tracking task

- Stories with similarity above a threshold $thresh_{yes/no}$ to the training story are marked YES



the tracking task

- misses and false alarms



$$P_{fa} =$$

$$P_{miss} =$$

tracking task - adaptation

- Consider that $sim(T_j, S_4) > thresh_{adapt}$

tracking task - adaptation

- add story S_4 to topic T_j and recompute model
- Danger of adapting with a false alarm story

tracking task - adaptation

- Adaptation
  - If $sim(T_j, S_i) > thresh_{yes/no}$ then story S_i is on topic T_j
  - If $sim(T_j, S_i) > thresh_{adapt}$ add story S_i to topic T_j and recompute model
  - $thresh_{adapt} > thresh_{yes/no}$

vector space for tracking

- Treat stories as "bags of words"
- Really as a vector of weighted features
  - Features are word stems (no stopwords)
  - Weights are a variant of tf-idf

IDF is incremental or retrospective

$$S = s_1 \ldots s_{|V|}$$

vector space for tracking

- Compare vectors by cosine of angle between the story and the topic.
  - If use same words in same proportion,stories are the same
  - If have no words in common, are about different topics

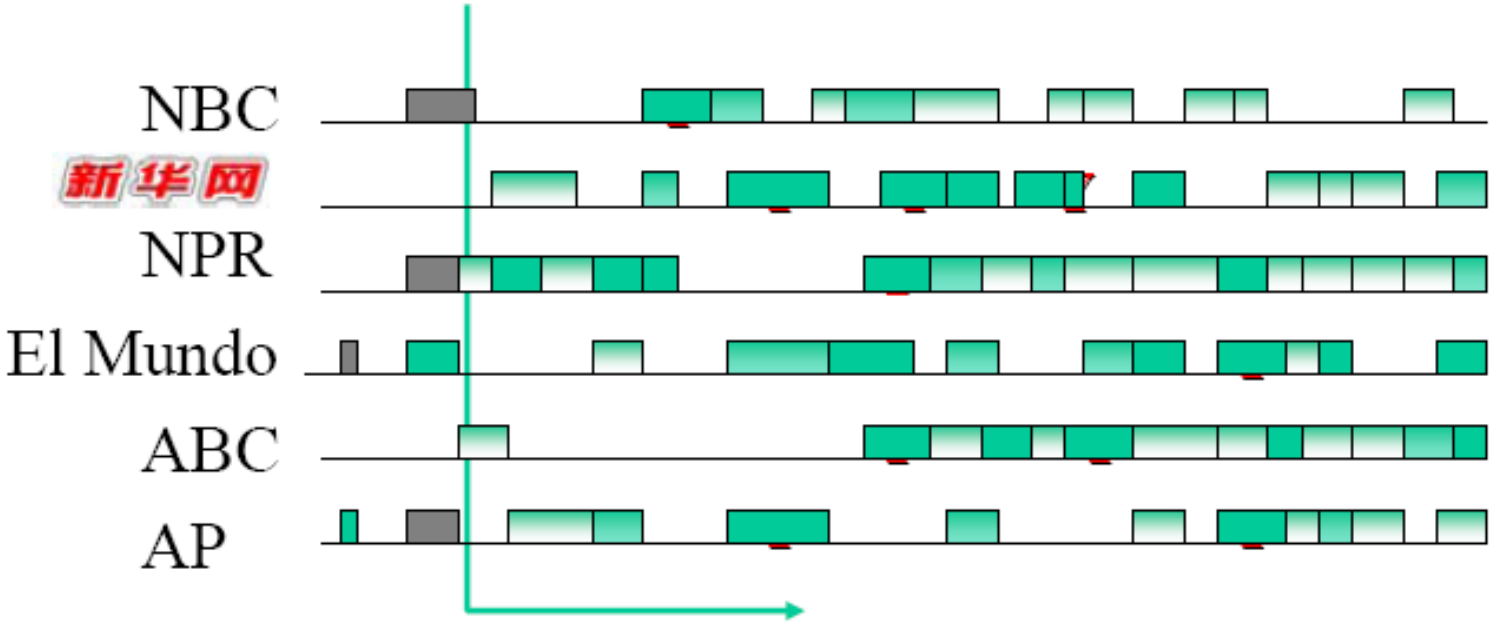
$$sim(S,T) = \frac{\sum_w s_w t_w}{\sqrt{\sum_w s_w^2 \sum_w t_w^2}}$$

measuring progress in TDT

- All tasks viewed as detection tasks (yes/no)
  - Is there a story boundary here?
  - Is this story on the topic being tracked?
  - Are these two stories on the same topic?
- Evaluations based on miss and false alarm
- Use linear combination as cost function

Evaluating tracking

- Perfect tracker says YES to on-topic stories and no to all other stories
- In reality, system emits confidence of topic

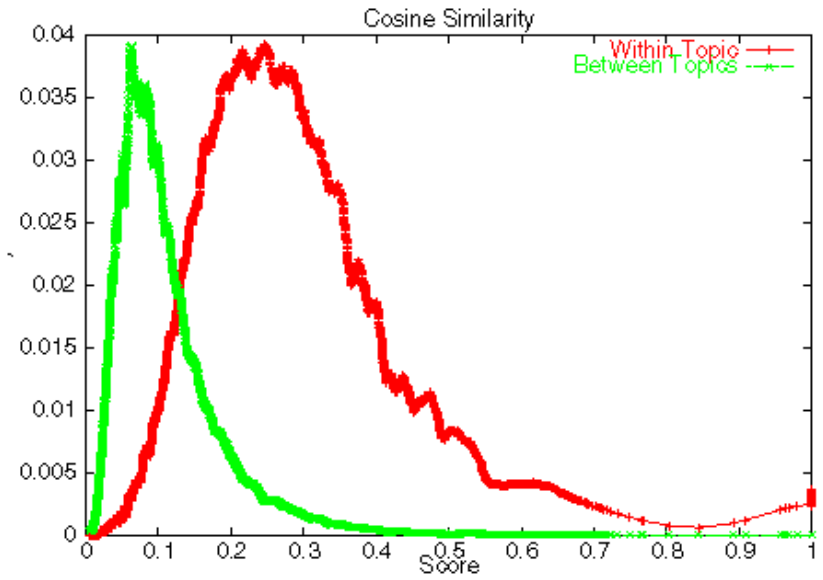


evaluating tracking

- At every score, there is a miss and false alarm rate
  - Any on-topic stories below score are misses
  - Any off-topic stories above score are false alarms

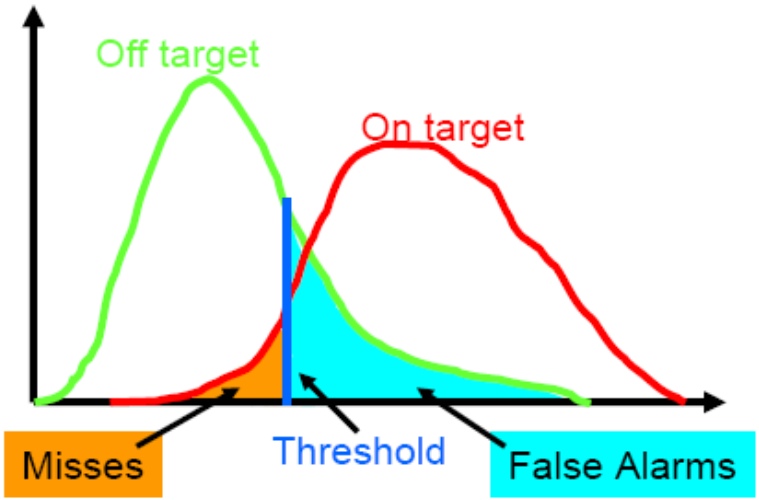
- Plot (false alarm, miss) pairs for every score
  - Result is an ROC curve
  - TDT uses a modification called the "DET curve" or "DET plot"

DET plots



- Green curve on left is "no"
- Red curve on right is "yes"
- X axis represents scores

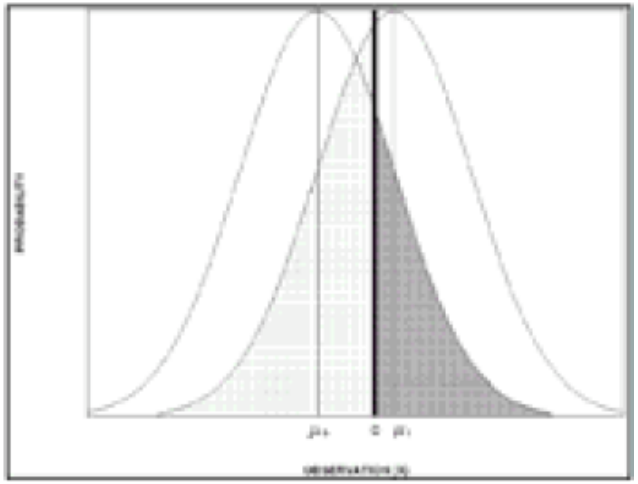
- Sweep through scores
- Note P(miss) and P(fa)
- Plot values at every score
- Plot of distribution of scores

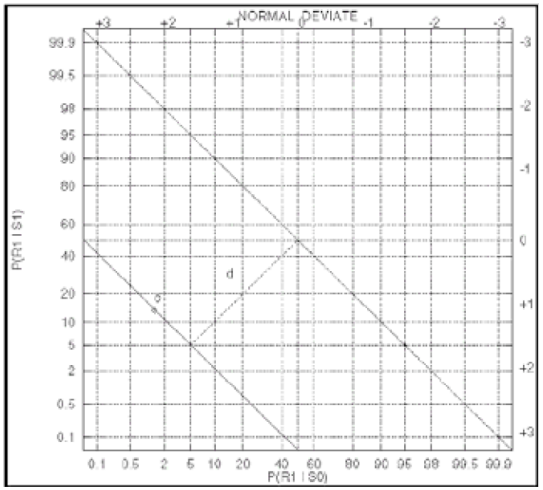


normal deviate ?

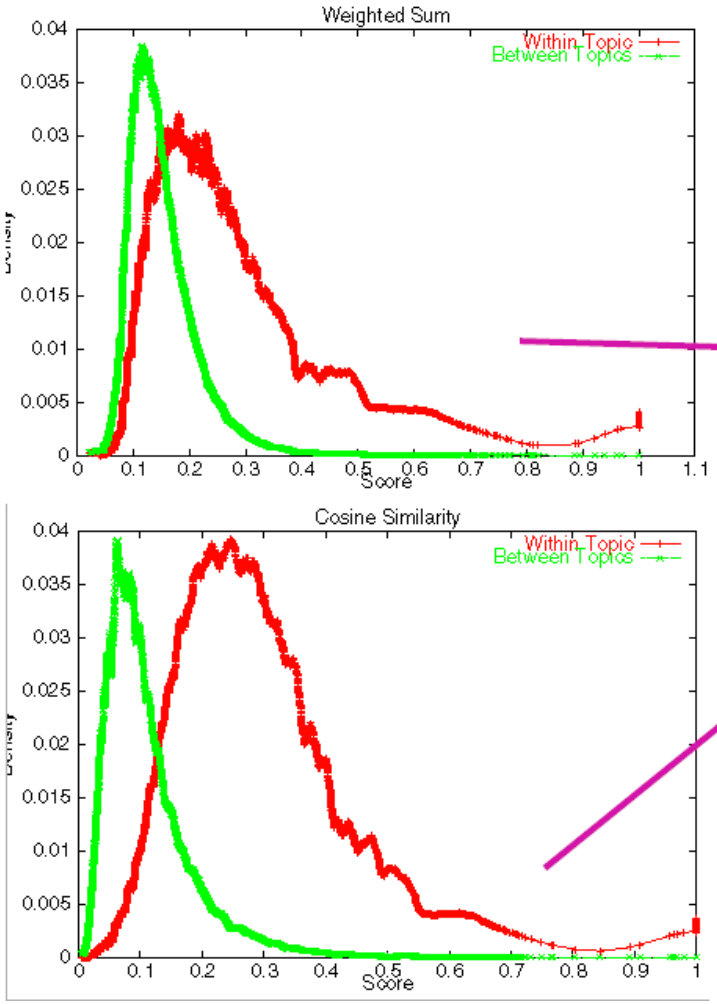
- Assume scores normally distributed with means μ_0 and μ_1
- Replace score with normal deviation
  - Normal distributions end up as straight lines
  - Intercept set by spread
  - Slope set by variance
- If $\mu_0=\mu_1$ then miss and false alarms in sync
  - Random performance
- Separation, $d=(\mu_1-\mu_0)0.5$

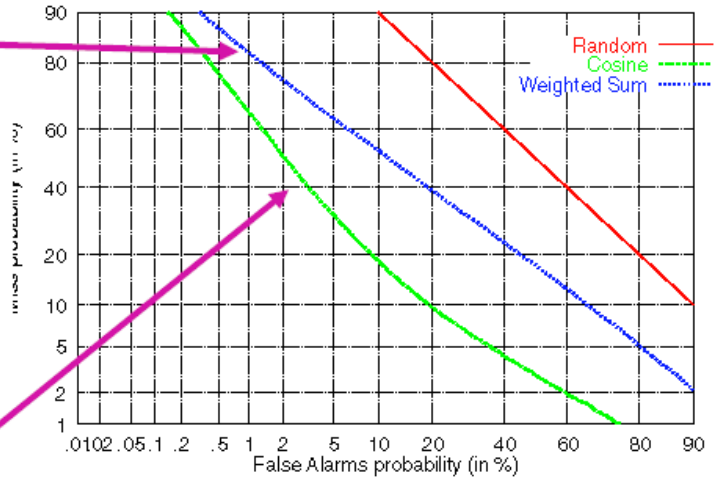
- Graphs from Martin et al, "The DET curve in assessment of detection task performance," Eurospeech, 1997.



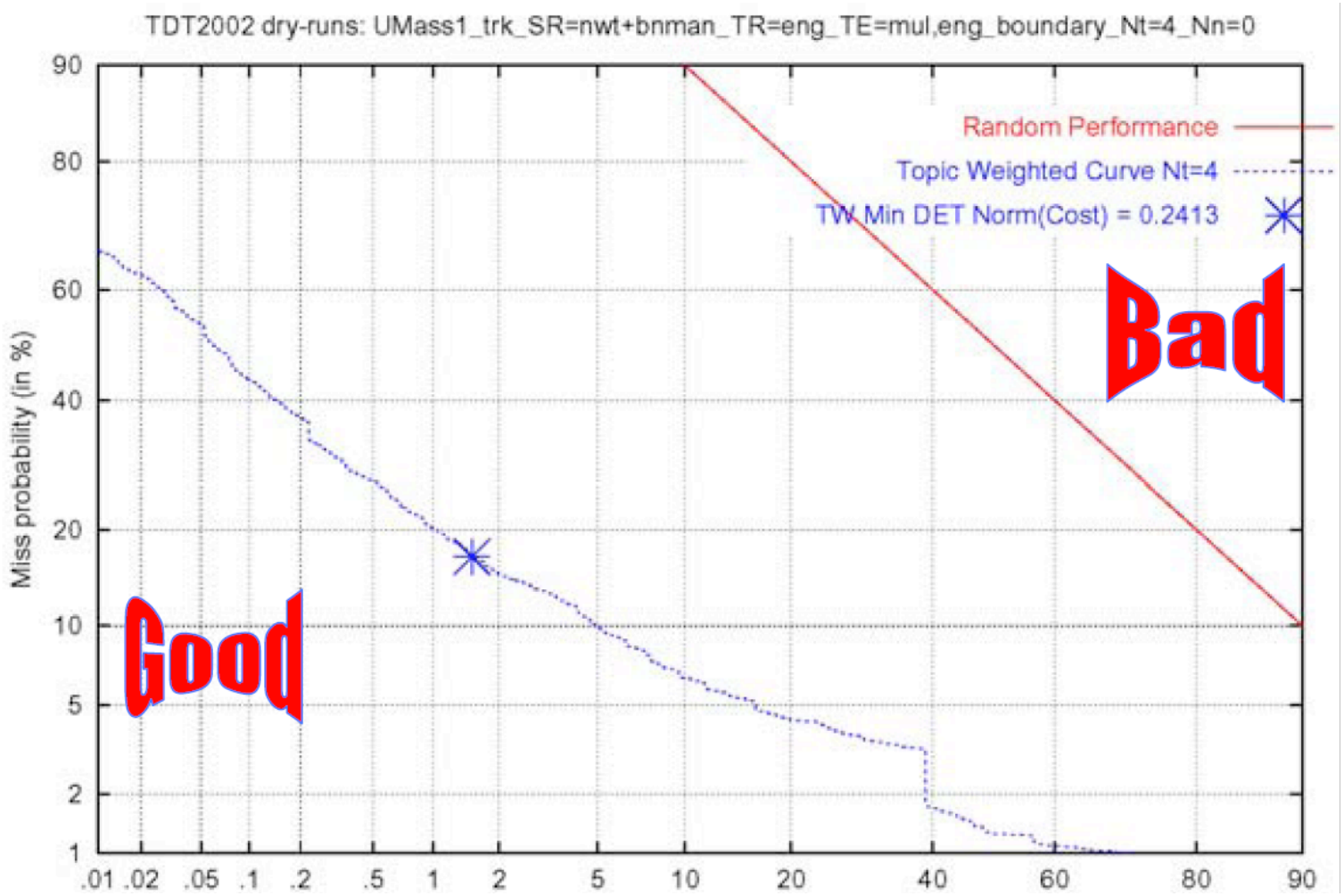


DET plot


Weighted Sum
Within Topic
Between Topics
Score
Cosine Similarity
Within Topic
Between Topics
Score


Random
Cosine
Weighted Sum
False Alarms probability (in %)

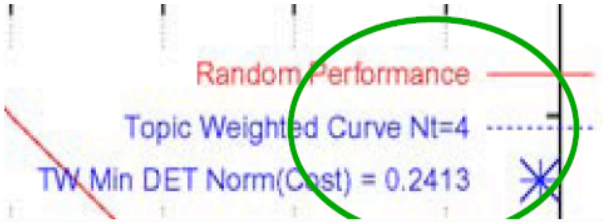
tracking DET curve (umass)



cost function

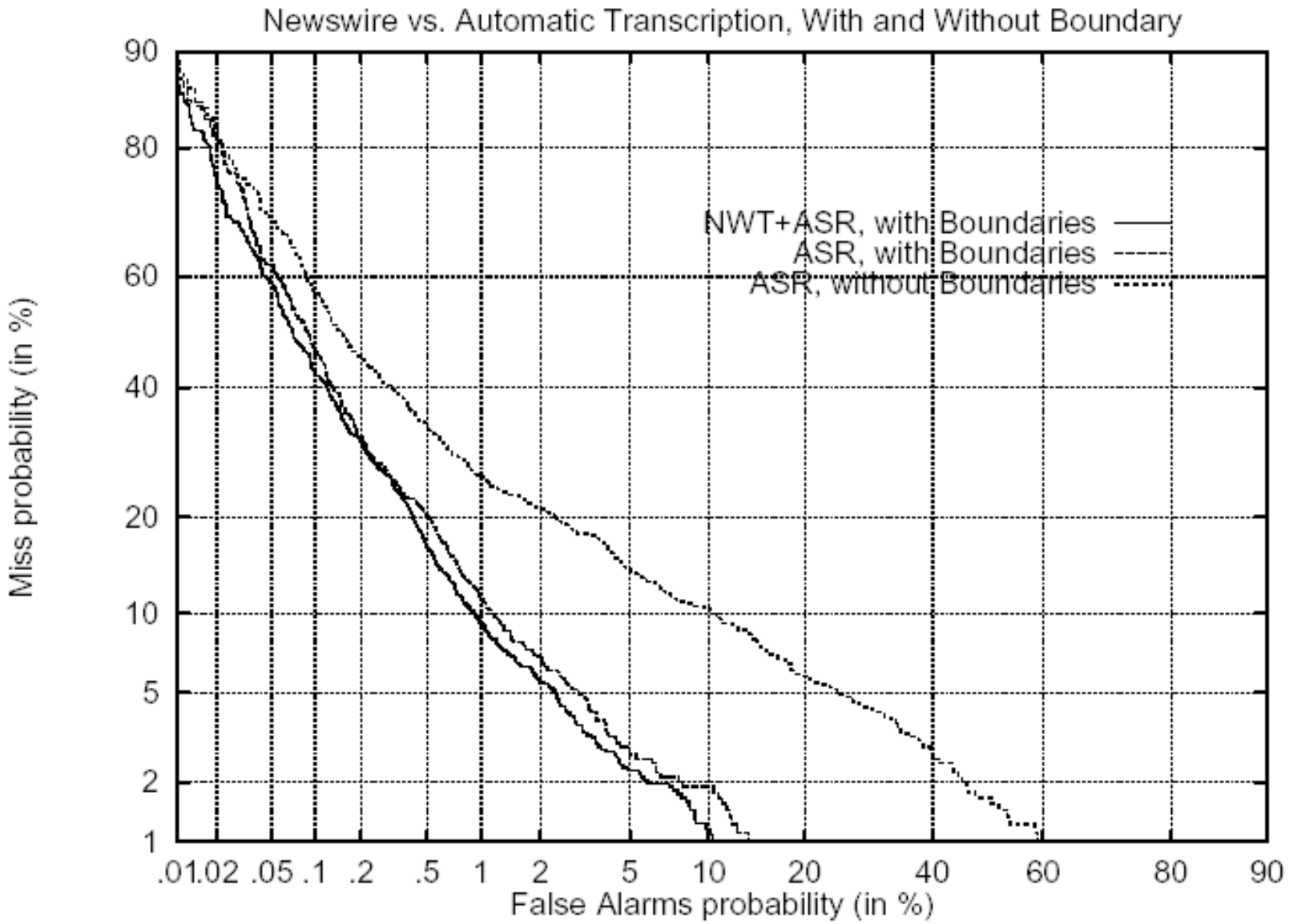
- Systems must choose "hard" decision point
  - Score that optimizes system performance
  - Determines a miss and false alarm pair
- Measure by cost (e.g., "tracking cost")

$$C_{track} = C_{miss} \cdot P_{miss} \cdot P_{target} + C_{fa} \cdot P_{fa} \cdot (1 - P_{target})$$



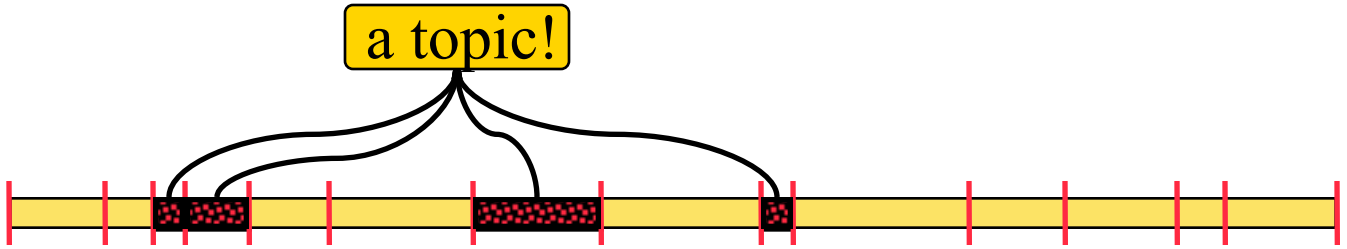
$$(C_{track})_{norm} = C_{track} \div min \left\{ \begin{array}{l} C_{track, P_{miss}=1, P_{fa}=0} \\ C_{track, P_{miss}=0, P_{fa}=1}) \end{array} \right\}$$

TDT Topic Tracking



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

- 3 Language Conditions:

English only
Mandarin only
English and Mandarin together

- 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

- Decision Deferral Conditions:

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

- 2 Story Boundary Conditions:

Reference story boundaries provided
No story boundaries provided

TDT summary

- Five technology evaluation tasks
  - Story segmentation – find story boundaries in broadcast news
  - Topic tracking – given sample stories, find rest on same topic
  - First story detection – detect onset of new event in the news
  - Cluster detection – group stories into events (unsupervised)
  - Story link detection – decide if two stories discuss same event
- Tracking and detection on event-based topics
  - Though most approaches are the same as those used for subject-based tasks
- All tasks are on-line (not batch) evaluations
  - Cluster detection task has a “retrospective” variation

outline

- news filtering
- TDT
- advanced TDT
- novelty detection

more realistic topic tracking [Leuski, Allan]

- Unrealistic assumptions about the user's behavior.
  - TREC filtering : forces the user to judge every document it labels relevant.
  - TDT tracking : avoids any dialog with the user.
- Intermediate scenario where the system may request the user's judgments for some of the returned documents but it does not ask her to judge all of them.
- Also the user may ignore some of the documents requested by the system.

more realistic topic tracking

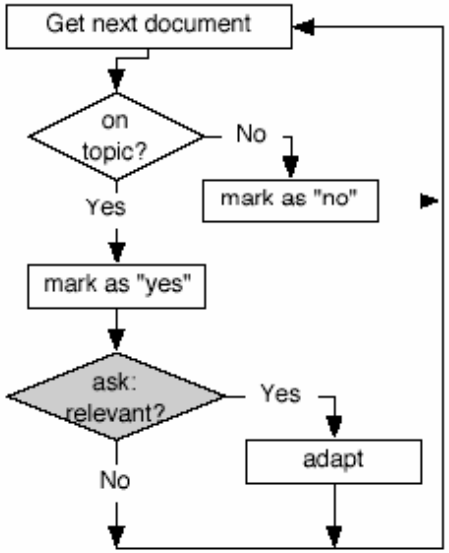


Figure 1: Shows the flow of control in TREC filtering task.

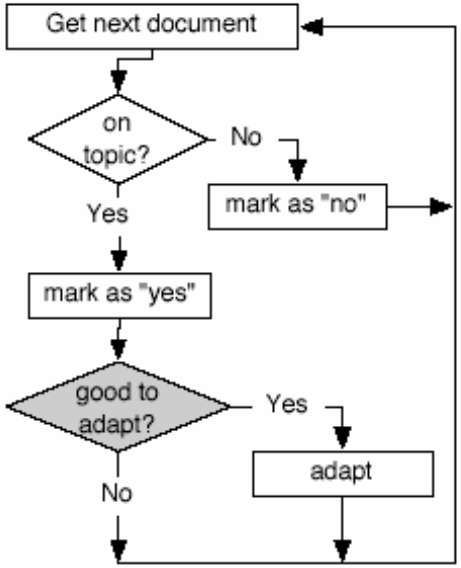


Figure 2: Shows the flow of control in TDT tracking task.

modeling interactive tracking

- System decides whether to involve the user and check if the user is interested in making the judgment

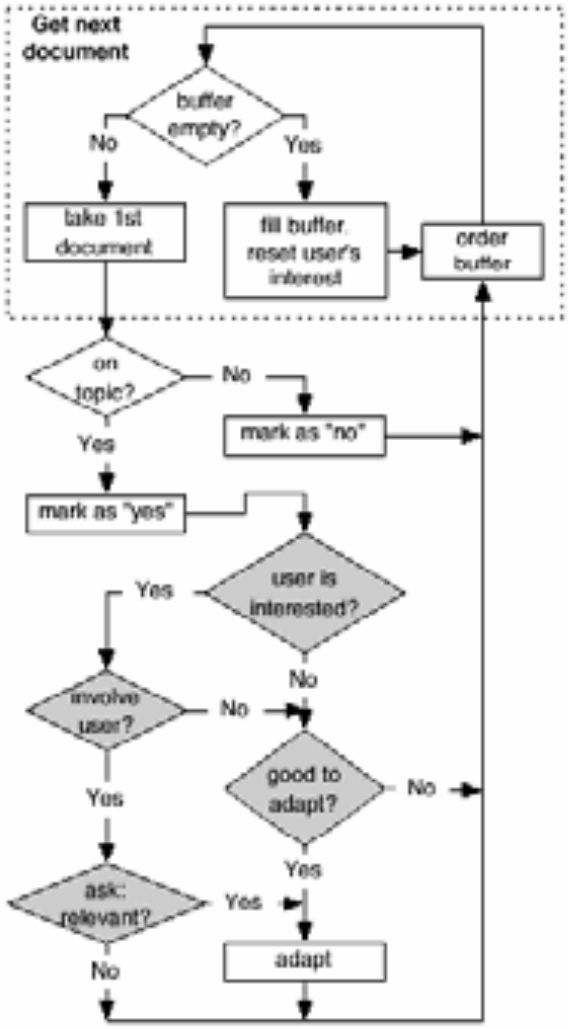


Figure 3: Shows the flow of control in the Interactive Tracking task.

modeling interactive tracking

- Consider a situation where the user interacts with the system at discrete time intervals.
  - At the beginning of each session the system lists the documents in the buffer for the user.
- Assumption is that the user will start at the top to the list and follow it down. the order of the documents in the session buffer is very important.
  - After it receives the user's feedback it adapts the topic representation and possibly re-orders the rest of the list

evaluate interactive tracking

$$C = C_{miss} \cdot P_{miss} \cdot P_{\text{target}} + C_{fa} \cdot P_{fa} \cdot (1 - P_{\text{target}})$$

- Where C_{miss} and C_{fa} are the costs of a missand false alarm.
- P_{miss} is the conditional probability of a miss.
- P_{fa} is the condition probability of a false alarm.
- P_{target} is the priori target probability.(1) target

evaluate interactive tracking

- In this paper, we compute the normalized version of the cost
- measure $C_{ost} = C/Cmin$,

$$C_{\min} = \min(C_{miss} \cdot P_{\text{target}}, C_{fa} \cdot (1 - P_{\text{target}})$$

- where

$$C_{miss}=1,\ C_{fa}=10,\ \text{and}\ P_{target}=0.02.$$

evaluate interactive tracking

- 4 types of documents:
  - the presented documents are labeled as relevant by the system.
  - the examined documents are the ones that the user reads.
  - the judged documents are the documents that user labeled as relevant or non-relevant for the system.
- TREC filtering: presented=examined=judged.
- TDT tracking:
  - presented = examined
  - judged documents=ø

evaluate interactive tracking

$$Activity = \frac{\#\text{ of judged}}{\#\text{ of examined}}$$

Does a decrease in activity result in decrease in performance?

evaluate interactive tracking

- Defer its labeling decisions until the user's next session.
- When the user begins the session, the system orders the documents in the buffer and presents them to the user.
- The user's goal is to locate the relevantmaterial in the news stream as quickly aspossible.
  - i.e relevant documents at the top of the list.
- Feedback should be as early as possible.
- i.e. close to the top of the list

evaluate interactive tracking conclusion

- Highlighted the simplifications that were made for evaluation purposes in the TREC and TDT news filtering tasks.
- A more realistic model of interaction.
- Smaller amount of necessary judgments from a user with only a small cost penalty.
- Documents may be“batched up” with no significant impact on cost, although the precision of the batch improves when the batches are smaller.
- Optimal cost : q, y+, y.
- If the user stops providing necessary judgments early, the cost of the final output is noticeably higher.
- However, if the user examines enough of the batch to be confident that the remainder is non-relevant, the cost stays low.

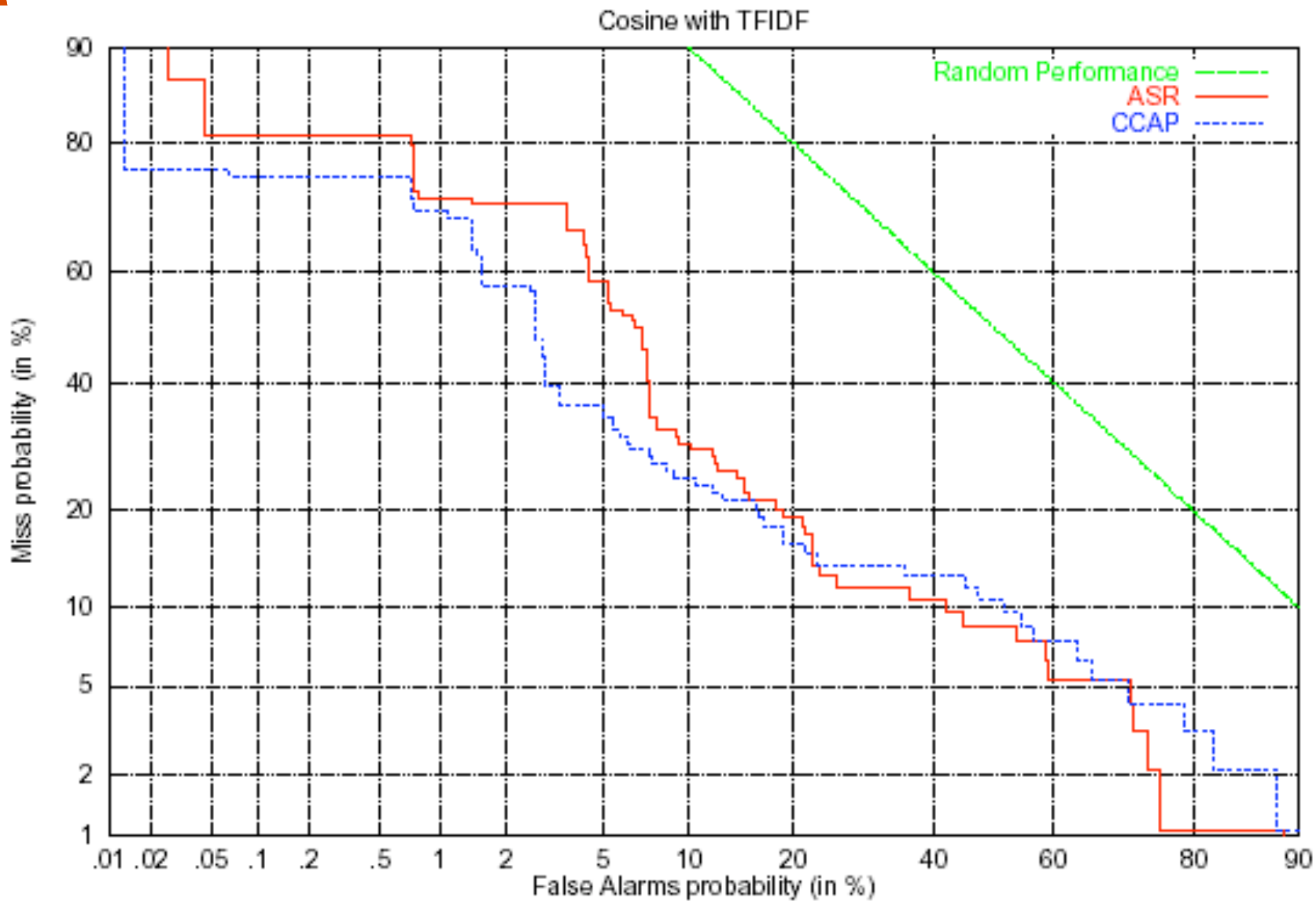
Exploration Exploitation Trade-off

- The importance or usefulness of negative feedback.
- Filtering is essentially online classification. But is a high precision task.
- Satisfy the users immediate need-- exploitation.
- All methods discussed this far cared only about exploitation

outline

- news filtering
- TDT
- advanced TDT
- novelty detection

TDT Novelty Detection



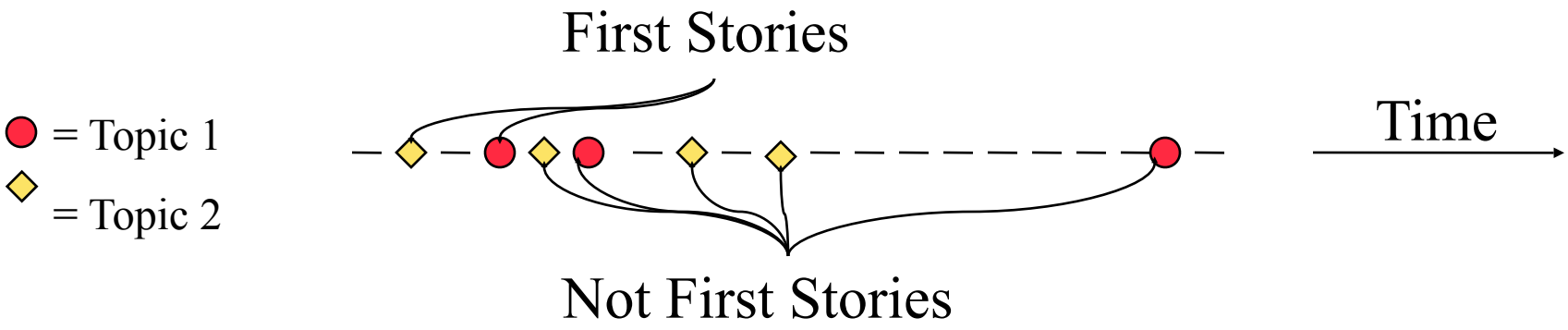
The First-Story Detection Task:

The First-Story Detection Task:

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

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)

First-Story Detection Conditions

- 1 Language Condition: **English only**
- 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

- Decision Deferral Conditions:

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

- 2 Story Boundary Conditions:

Reference story boundaries provided
No story boundaries provided

novelty detection

- Novelty Detection Approaches
  - VSM + Clustering Techniques
  - Learning with multiple features
  - Support Vector Machines
  - Non-parametric Bayesian method
- TDT 2004 Evaluation Results

VSM + Clustering Techniques

- Vector Space Model (traditional IR technique)
  - Documents are represented as vectors
  - TFIDF term weighting is used, and similarity measure is chosen (e.g. cosine)

- Clustering
  - Lookahead window: GAC clustering
    - Slight improvement (appr. 3%)
  - Past window: incremental clustering
    - No improvement

Supervised learning with informative features

- Convert novelty detection to a supervised learning task
  - Positive/Negative data: novel/non-novel stories
- Build the model
  - Choose a learning algorithm (Logistic regression, SVM, etc.)
  - Try to select good features (features we tried: cosine score, cluster size, time stamp)
- Gives better story-weighted results in the past evaluations

Support Region Estimation with SVM

- Treat the problem as “density estimation”
- Use one-class Support Vector Machines
  - One of the best performing supervised learning algorithms
  - Suitable high dimensional, sparse data
  - Has been successfully applied to novelty detection in hand-written digits
- Performance in novelty detection task
  - Worse than “VSM + clustering”
  - Unsupervised kernel selection is hard

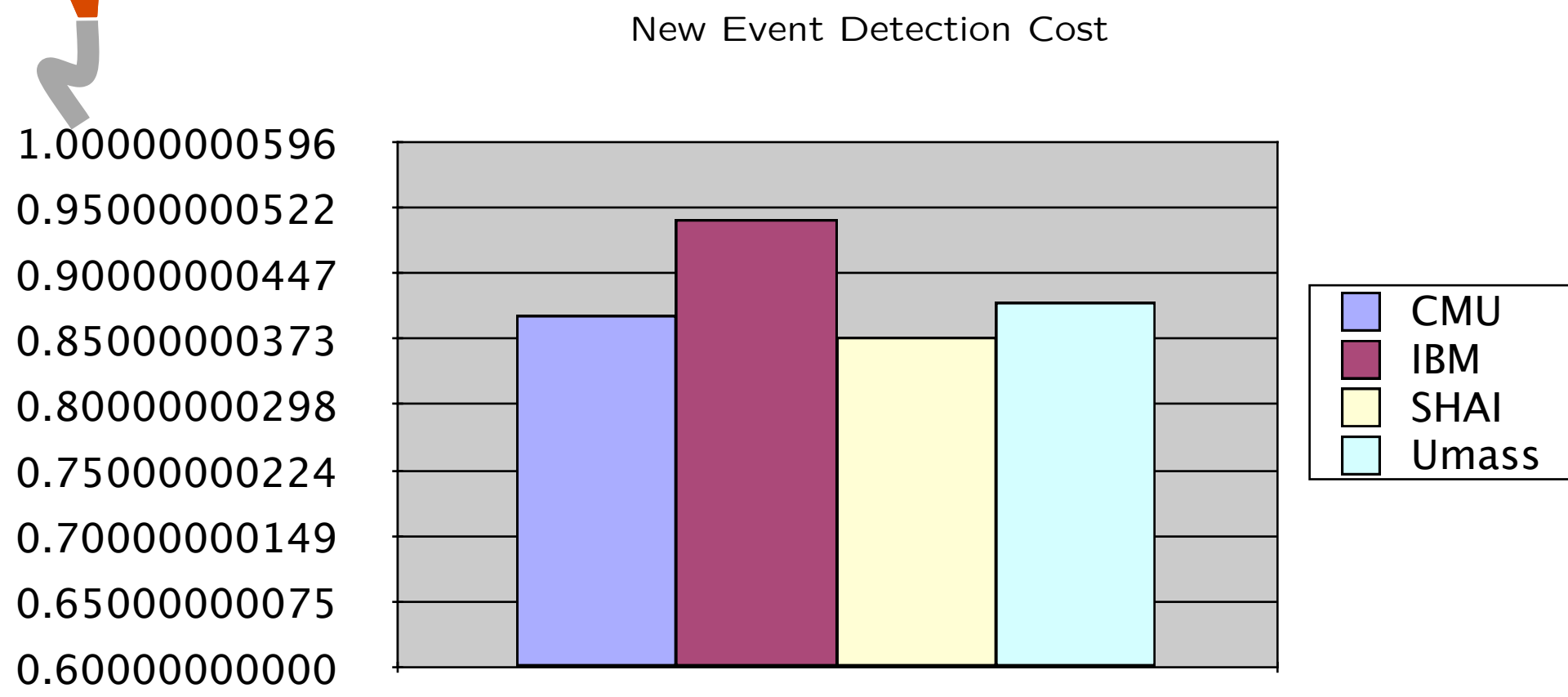
Non-parametric Bayesian method (Dirichlet Process Mixture Model)

- Density estimation method in statistics
  - Converges to empirical distribution asymptotically
  - Recently has been applied in machine learning/bio-informatics community
- Advantages:
  - Handle increasing number of clusters
  - Probabilistic interpretation
- Performance:
  - Comparable to "VSM+clustering"
  - More expensive

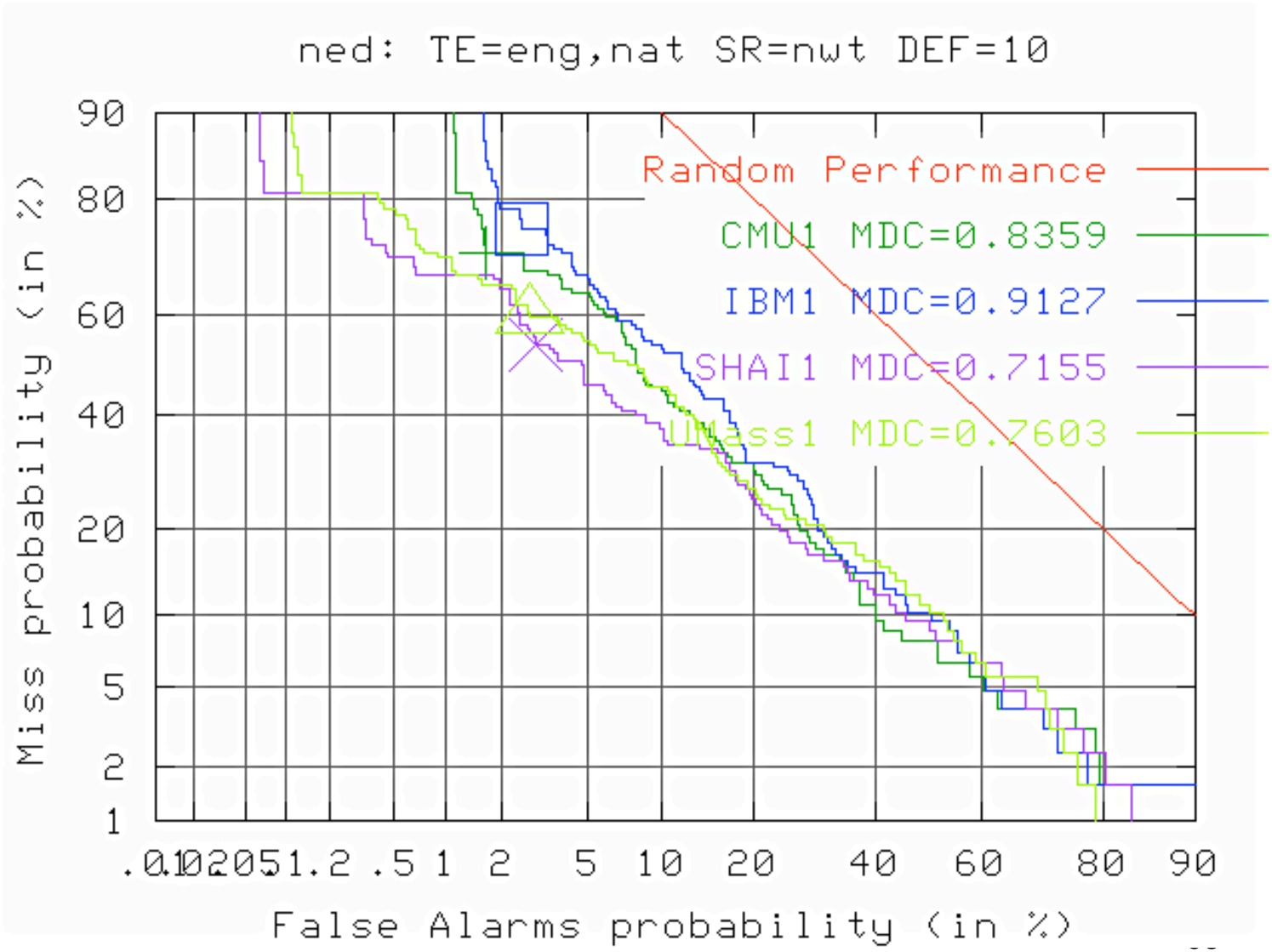
TDT 2004 NED Task

- Dataset
  - Large size: around 280k documents
  - Time period: Apr. 2003 – Sep. 2003 (6 months)
- Submitted Method:
  - VSM + GACINCR clustering
- System parameter tuning:
  - Use TDT3 corpus

TDT 2004 NED Results



TDT 2004 NED DET-Curve

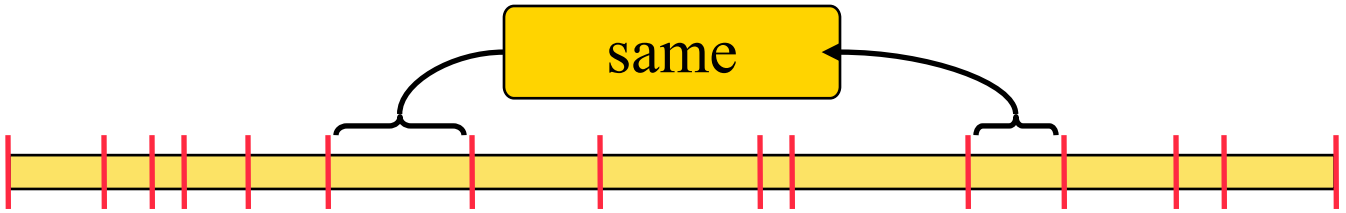


References

- CMU TDT 2001/2002/2003 report.
- Yiming Yang, Tom Pierce and Jaime Carbonell. A Study on Retrospective and Online Event Detection. SIGIR 1998.
- Yiming Yang, Jian Zhang, Jaime Carbonell and Chun Jin. Topic-conditioned Novelty Detection. SIGKDD 2003.
- Jian Zhang, Yiming Yang and Jaime Carbonell. New Event Detection with Nearest Neighbor, Support Vector Machines and Kernel Regression. CMU Tech. Report CMU-CS-04-118 (CMU-LTI-04-180).
- Jian Zhang, Zoubin Ghahramani and Yiming Yang. A Probabilistic Model for Online Document Clustering with Applications to Novelty Detection. NIPS 2004.

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.

Link Detection Conditions

- 1 Language Condition: **English only**
- 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

- Decision Deferral Conditions:

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

- 1 Story Boundary Condition:

Reference story boundaries provided

Example Performance Measures: