news filtering
topic detection and tracking
outline

- news filtering
- TDT
- advanced TDT
- novelty detection
Google alerts
RSS feeds


- 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
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
    • 3 Mandarin
    ABC CNN
    PRI VOA
    NBC MNB
    APW NYT
    VOA XIN ZBN
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.
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.*

Transcription: text (words) →

- **Story:**
- **Non-story:**

(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:

<table>
<thead>
<tr>
<th>Maximum Decision Deferral Period</th>
<th>Text</th>
<th>Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (words)</td>
<td>Mandarin (characters)</td>
<td>English &amp; Mandarin (seconds)</td>
</tr>
<tr>
<td>100</td>
<td>150</td>
<td>30</td>
</tr>
<tr>
<td>1,000</td>
<td>1,500</td>
<td>300</td>
</tr>
<tr>
<td><strong>10,000</strong></td>
<td><strong>15,000</strong></td>
<td><strong>3,000</strong></td>
</tr>
</tbody>
</table>
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

NBC

NPR

(1) El Mundo

ABC

AP
to this

NBC

新华网 (Xinhua)

NPR

El Mundo

ABC

AP
TDT data

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

<table>
<thead>
<tr>
<th>English</th>
<th>Mandarin</th>
<th>MT</th>
<th>Nat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign</td>
<td>Arabic</td>
<td>MT</td>
<td>Nat</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Print</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>ASR</td>
</tr>
<tr>
<td></td>
<td>Manual</td>
</tr>
<tr>
<td></td>
<td>Mandarin</td>
</tr>
<tr>
<td></td>
<td>Arabic</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

New This Year: not guaranteed to be off-topic
Topic Tracking Conditions

9 Training Conditions:

<table>
<thead>
<tr>
<th>Training Language</th>
<th>English</th>
<th>Mandarin</th>
<th>Both Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_t$</td>
<td>1 (E)</td>
<td>1 (M)</td>
<td>1 (E), 1(M)</td>
</tr>
<tr>
<td>English (E)</td>
<td>2 (E)</td>
<td>2 (M)</td>
<td>2 (E), 2(M)</td>
</tr>
<tr>
<td>Mandarin (M)</td>
<td>4 (E)</td>
<td>4 (M)</td>
<td>4 (E), 4(M)</td>
</tr>
</tbody>
</table>

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
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.
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 $\text{thresh}_{\text{yes}/\text{no}}$ to the training story are marked YES
the tracking task

- misses and false alarms

Truth: N N Y Y

\[
P_{fa} = \frac{\text{blue}}{\text{gray}}
\]

\[
P_{miss} = \frac{\text{yellow}}{\text{green}}
\]
tracking task - adaptation

- Consider that $\text{sim}(T_j, S_4) > \text{thresh}_{\text{adapt}}$
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 \( \text{sim}(T_j, S_i) > \text{thresh}_{\text{yes/no}} \) then story \( S_i \) is on topic \( T_j \)
- If \( \text{sim}(T_j, S_i) > \text{thresh}_{\text{adapt}} \) add story \( S_i \) to topic \( T_j \) and recompute model
- \( \text{thresh}_{\text{adapt}} > \text{thresh}_{\text{yes/no}} \)
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

\[ S = s_1 \ldots s_{|V|} \]

IDF is incremental or retrospective
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

- Sweep through scores
- Note $P(\text{miss})$ and $P(\text{fa})$
- Plot values at every score
- Plot of distribution of scores

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

Cosine Similarity

Off target
On target
Threshold
Misses
False Alarms
normal deviate?

- Assume scores normally distributed with means \( \mu_0 \) and \( \mu_1 \)
- 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, \( d = (\mu_1 - \mu_0)^{0.5} \)

DET plot
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”)

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

\[
(C_{\text{track}})_{\text{norm}} = \frac{C_{\text{track}}}{\min \left\{ C_{\text{track}}, P_{\text{miss}} = 1, P_{\text{fa}} = 0 \right\}, C_{\text{track}}, P_{\text{miss}} = 0, P_{\text{fa}} = 1}
\]
TDT Topic Tracking

Newswire vs. Automatic Transcription, With and Without Boundary

Miss probability (in %) vs. False Alarms probability (in %)

- NWT+ASR, with Boundaries
- ASR, with Boundaries
- ASR, without Boundaries
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

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

- Where \( C_{\text{miss}} \) and \( C_{\text{fa}} \) are the costs of a miss and false alarm.
- \( P_{\text{miss}} \) is the conditional probability of a miss.
- \( P_{\text{fa}} \) is the condition probability of a false alarm.
- \( P_{\text{target}} \) is the priori target probability.
evaluate interactive tracking

• In this paper, we compute the normalized version of the cost measure $C_{\text{cost}} = C / C_{\text{min}}$,

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

• where

$$C_{\text{miss}} = 1, C_{fa} = 10, \text{ and } P_{\text{target}} = 0.02.$$
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 relevant material in the news stream as quickly as possible.  
  - 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

Cosine with TFIDF

Miss probability (in %)

False Alarms probability (in %)

Random Performance
ASR
CCAP
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.*

- First Stories
  - \( \bullet \) = Topic 1
  - \( \diamond \) = Topic 2

- Not First Stories

- There is no supervised topic training (like Topic Detection)
First-Story Detection Conditions

1 Language Condition:  

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:

<table>
<thead>
<tr>
<th>Maximum decision deferral period in # of source files</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

2 Story Boundary Conditions:

<table>
<thead>
<tr>
<th>Reference story boundaries provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>No story boundaries provided</td>
</tr>
</tbody>
</table>
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

- **Submitted Method:**
  - VSM + GACINCR clustering

- **System parameter tuning:**
  - Use TDT3 corpus
TDT 2004 NED Results

New Event Detection Cost

<table>
<thead>
<tr>
<th>CMU</th>
<th>IBM</th>
<th>SHAI</th>
<th>Umass</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.60000000000</td>
<td>0.65000000075</td>
<td>0.70000000149</td>
<td>0.75000000224</td>
</tr>
<tr>
<td>0.80000000298</td>
<td>0.85000000373</td>
<td>0.90000000447</td>
<td>0.95000000522</td>
</tr>
<tr>
<td>1.00000000596</td>
<td>0.75000000224</td>
<td>0.90000000447</td>
<td>0.95000000522</td>
</tr>
<tr>
<td>0.60000000000</td>
<td>0.65000000075</td>
<td>0.70000000149</td>
<td>0.75000000224</td>
</tr>
<tr>
<td>0.80000000298</td>
<td>0.85000000373</td>
<td>0.90000000447</td>
<td>0.95000000522</td>
</tr>
<tr>
<td>1.00000000596</td>
<td>0.75000000224</td>
<td>0.90000000447</td>
<td>0.95000000522</td>
</tr>
</tbody>
</table>
TDT 2004 NED DET-Curve

ned: TE=eng,nat SR=nwt DEF=10

Random Performance
CMU1 MDC=0.8359
IBM1 MDC=0.9127
SHAI1 MDC=0.7155
UMass1 MDC=0.7603

Miss probability (in %)

False Alarms probability (in %)
References

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 $N_t = 1$. 

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:**

<table>
<thead>
<tr>
<th>Maximum decision deferral period in # of source files</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>100</td>
</tr>
</tbody>
</table>

- **1 Story Boundary Condition:**

  Reference story boundaries provided
Example Performance Measures:

Tracking Results on Newswire Text (BBN)

- Random Performance
- BBN1: NWT English
- BBN1: NWT Mandarin

Graph showing miss probability vs. false alarms probability for English and Mandarin languages.

Bar chart showing normalized tracking cost for English and Mandarin.