Chapter 10: Topic Detection and Tracking (TDT)

Some slides from “Overview NIST Topic Detection and Tracking
-Introduction and Overview” by G. Doddington
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: 30,000 stories
    - Mandarin: 10,000 stories
  - 11 Different Sources:
    - English: 8
    - Mandarin: 3
    - ABC, CNN, VOA, PRI, VOA, XIN, NBC, MNB, ZBN, APW, NYT

* see [http://www.itl.nist.gov/iaui/894.01/tdt3/tdt3.htm](http://www.itl.nist.gov/iaui/894.01/tdt3/tdt3.htm) for details
† see [http://morph.ldc.upenn.edu/Projects/TDT3/](http://morph.ldc.upenn.edu/Projects/TDT3/) for details
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)
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.

training data

not guaranteed to be off-topic

test data

on-topic
unknown
unknown
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:

<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>
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 $N_t = 1$. 
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:
  \[ C_{Det} = C_{Miss} \cdot P_{Miss} + C_{FA} \cdot P_{FA} \]
  (e.g. \( C_{Miss} = 0.2 \), \( C_{FA} = 0.98 \))
- Detection Cost is normalized to lie between 0 and 1:
  \[ (C_{Det})_{Norm} = \frac{C_{Det}}{\min\{C_{Miss}, C_{FA}\}} \]
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

Actual Decision Cost
Minimum DET Graph Cost

Normalized Tracking Cost

 normalized cost for BBN1, CMU1, Dragon1, UPenn1, GE1, Ulowa1, U Md1
1999 TDT3 Tracking Results

Required Evaluation Condition
Story Segmentation using Decision Trees

- **tokenizer** → **sentence detection** → **POS tagger**
  - words → **morph table** → **morphs**
  - tags → **feature extraction** → **features**
  - P(seg|context) → **decision tree scoring** → **peak extraction** → **refinement**

**Decision Tree**

- silence length$>0.56s$
  - yes: P(seg) = 7%
  - no: P(seg) = 16.5%

- new nouns $>7$
  - yes: P(seg) = 2.6%
  - no: 5%

- right key bigram distance $<=1$
  - yes: 7%
  - no: 30%

- silence length$>0.32s$
  - yes: 5.3%
  - no: 19%

- right key bigram distance $<=15$
  - yes: 25%
  - no: 55%

---

**Story Segmentation and Topic Detection in the Broadcast News Domain**

S. Dhareniyangoda  M. Ponz  J.S. McCreary  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

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

\[ P(q_{N+1} | q_1...q_N) = \sum_{i=1}^{k} P(q_{N+1} | M_i) P(M_i | q_1...q_N) \]
Using Relevance Models in Link Detection

• Question:
  • Are stories $S_1$ and $S_2$ linked?

• Approach
  • Create a 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

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

\[
P(w \mid S_1) = P(w \mid q_1...q_N)
\]

\[
= \sum_{D \in R} P(w \mid D)P(D \mid q_1...q_N)
\]
Measuring Distances

Kullback-Leibler Distance

\[ D(S_1 \parallel S_2) = \sum_w P(w \mid S_1) \log \frac{P(w \mid S_1)}{P(w \mid S_2)} \]

Symmetric Kullback-Leibler Distance

\[ D_{sym}(S_1 \parallel S_2) = \frac{1}{2} \left( D(S_1 \parallel S_2) + D(S_2 \parallel S_1) \right) \]

Kullback-Leibler Distance with “Clarity”

\[ D_{Cl}(S_1 \parallel S_2) = \sum_w P(w \mid S_1) \log \frac{P(w \mid S_2)}{P(w \mid GE)} \]

(GE : general english)
Comparison on Training Data

![Comparison on Training Data Graph](image)

- **Best Cosine + TF.IDF** (min. cost: 0.09)
- **Best Relevance Model** (min. cost: 0.07)
Comparison on Evaluation Data

Random
Cosine (TF.IDF) (min: 0.27)
Relevance Model (min: 0.24)
Confidence Intervals
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