# Presentation “TimeMines: Constructing Timelines with Statistical Model of Word Usage”

Tao Xia

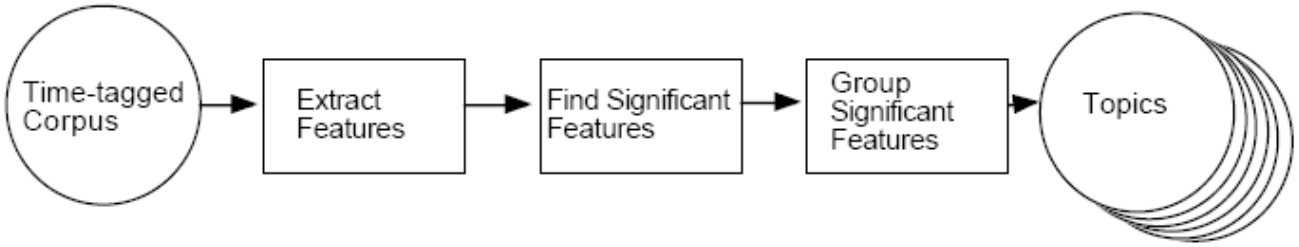
November 23, 2006

# Motivation

- TimeMines displays important topics in the corpus, and their coverage and time spans.
- Detects, ranks, and groups semantic features based on their statistical properties.

# System Overview

- Extracting Features
- Finding Significant Features
- Grouping Significant Features
  - Default model is a stationary random model: the occurrence of a feature does not vary with time. The interesting ones are those features that violate the default model



# Extracting Features

- Extract noun phrases and named entities from text
  - Named entity: a person, location, organization
  - Noun phrases: matched regular expression (Noun|Adjective)*Noun
- All features extracted has time attached. We have a serious of the appearance of features over time.

# Finding Significant Features

- Contingency table

| | $f_0$ | $\overline{f_0}$ |
|---|---|---|
| $t \in t_0$ | a | b |
| $t \notin t_0$ | c | d |

- Perform x^2 test to determine the significant of the association.
- Under default model, the probability of a feature are the same over time. The results of the x^2 test are not significant under default model (significant level 5%)
- f0 is a feature, t0 is the time span

# Grouping Significant Features

- Many features may associated with one topic.
- Assumption: two features f0 and f1 have independent distributions implies that P(f0)=P(f0|f1)
- Contingency table

| | $f_j$ | $\overline{f_j}$ |
|---|---|---|
| $f_k$ | a | b |
| $\overline{f_k}$ | c | d |

# Evaluation

- Topic Detection and Tracking (TDT)
  - CNN broadcast news and Reuters newswire from July 1, 1994 to June 30, 1995
  - 15683 stories
- TDT-2
  - ABC News, CNN, Public Radio International, Voice of American, the New York Times, and the Associated Press newswire from January 1, 1998 to June 30, 1998
- JTAG – noun phrase
- Badger IE – named entity

# TDT-1

| Feature | Date Range |
|---|---|
| Oklahoma City (loc) | April 20 - April 29 |
| Kobe (loc) | Jan 16 - Jan 20 |
| Oklahoma (loc) | April 20 - April 27 |
| FBI (org) | April 20 - April 27 |
| Timothy McVeigh (pers) | April 21 - April 28 |
| NATO (org) | June 2 - June 5 |
| John Doe (pers) | April 21 - April 27 |
| Japan (loc) | Jan 16 - Jan 20 |
| Osaka (loc) | Jan 16 - Jan 18 |
| NATO (org) | May 25 - May 27 |

**Table 1: Top 10 named entities in TDT-1 by $\chi^2$ value**

| Feature | Date Range |
|---|---|
| oklahoma | April 20 - April 29 |
| oklahoma city | April 20 - April 29 |
| f-16 | June 2 - June 5 |
| kobe | Jan 16 - Jan 20 |
| bosnia | May 25 - June 8 |
| bombing | April 20 - April 29 |
| quake | Jan 16 - Jan 20 |
| bosnian serbs | May 25 - June 8 |
| serbs | May 24 - June 6 |
| bosnian | May 25 - May 26 |

**Table 2: Top 10 noun phrase features in TDT-1 by $\chi^2$ value**

| Story | Date Range |
|---|---|
| Oklahoma City Bombing | April 20 - April 29 |
| Earthquake in Kobe, Japan | Jan 16 - Jan 20 |
| F-16 shot down over Bosnia | June 2 - June 5 |
| NATO forces in Bosnia | May 25 - May 27 |
| Flooding in California | Jan 10 - Jan 11 |
| NATO forces in Bosnia | May 29 - May 31 |
| Senate debates Balanced Budget | Feb 28 - Mar 2 |
| Russia/US Summit | May 6 - May 10 |
| Two Americans Sentenced in Iraq | Mar 25 - Mar 27 |
| Henry Foster rejected by Senate as Surgeon General | June 21 - June 22 |

**Table 3: Top 10 stories as calculated by named entity statistics (labels manually assigned)**

# TDT-2

| Feature | Date Range |
|---|---|
| Iraq (loc) | Jan 26 - Feb 25 |
| Jonesboro (loc) | March 24 - March 31 |
| Iraqi Foreign Ministry (org) | Feb 21 - Feb 23 |
| Nagano (loc) | Jan 31 - Feb 23 |
| Davos (loc) | Jan 31 - Feb 2 |
| Barry Goldwater (pers) | May 29 - May 30 |
| Pol Pot (pers) | March 15 - March 19 |
| Ross Rebagliati (pers) | Feb 11 - Feb 14 |
| Duisenberg (pers) | May 2 - May 3 |
| v.o. (loc) | May 20 - May 22 |

**Table 5: Top 10 named entities in TDT-2 by $\chi^2$ value**

| Feature | Date Range |
|---|---|
| henshen | March 30 - March 31 |
| easter | May 9 - May 13 |
| three-hour meeting | Feb 22 - Feb 23 |
| arat | March 9 - March 11 |
| iraq | Jan 26 - Feb 25 |
| clock saturday | Feb 22 - Feb 23 |
| st patrick | March 15 - March 17 |
| jonesboro | March 24 - March 31 |
| naval armada | Feb 22 - Feb 23 |
| westphele | April 14 - April 15 |

**Table 6: Top 10 noun phrase features in TDT-2 by $\chi^2$ value**

| Topic | Date Range |
|---|---|
| U.S. Confrontation with Iraq<br>*iraq, iraqi foreign ministry, three-hour meeting, ...* | Jan 26–Feb 26 |
| Shooting at Westside Middle School, Jonesboro<br>*jonesboro, westside middle, westside middle school, ...* | March 24–April 1 |
| 1998 Winter Olympics<br>*nagano, ross rebagliati, medal, snowboarder, ...* | Jan 30–Feb 25 |
| Barry Goldwater dies<br>*barry goldwater, senator barry goldwater, arizona senator* | May 29–May 31 |
| Pol Pot dies<br>*pol pot, khmer rouge, killing fields, ...* | April 14–19 |
| Introduction of the Euro<br>*duisenberg, jean-claude trichet, european central bank, ...* | May 1–May 5 |
| Unrest in Indonesia<br>*habibie, indonesia, president suharto, ...* | May 12–May 27 |
| Crash of China Airlines A-300 Airbus | Feb 16–Feb 17 |

# Validation

- Randomization Test
  - To prove the patterns are valid
  - Ran TimeMines on shuffled corpus
  - Hypothesis: A simple statistical ranking of term occurrence and co-occurrence can identify and group relevant documents into coherent time-dependent stories.

# Validation

- Topic Based Evaluation
  - Use January 1996 Facts of File as truth
  - Misalignment between the corpus and the truth set
  - Hypothesis: These stories will prove comprehensible and useful to human users

# Feedback

- Positive
  - Use simple statistical model to retrieve and correlated significant features
- Negative
  - More detail about the x^2
    - SUM_table[(observed-expected)^2/expected]
  - There is no formula or any number on processing features (interesting to see the difference between the top features)
  - Need a better evaluation method