Language Technologies for e-Commerce

Lecture #4

March 23, 2001

Topic Detection and Tracking

Introduction

Topic Tracking and Detection is

1. the task of picking out "interesting" stories from a continuous stream of news (from one or more sources)

2. a project to advance the state of the art in the above task

What is an "Interesting" Story?

- one that provides new information about some topic/event the user wants to learn more about

- one that describes a new event about which the user has previously not known

Why do we need TDT?

- continuous flood of information, but only limited human attention

- finding new info is a key, recurring problem

TDT Project Tasks

- **Segmentation**
  Given a stream of words, find the story boundaries.
- **Detection**
  - **online**
    Assign a stream of news stories to appropriate categories.
  - **retrospective**
    Group a collection of previously-received news stories into sets of similar documents.
- **Tracking**
  Given a small number of stories about an "interesting" event, find all future stories covering that event.
- **First-Story Detection**
  From a stream of news stories, pick out those with the very first mention of any event.
- **Story Linking**
  Given two arbitrary news stories, determine whether or not they discuss the same event.

Differences between TDT and IR/TC

The main difference is that TDT considers its collections to be temporally ordered -- each document has an associated timestamp.

- incomplete information in early stories
- topic drift

Compared to traditional text categorization tasks, tracking has only a miniscule amount of training information available (at most four instances, compared to tens or hundreds for TC), and often only a very small number of instances to be found.

TDT uses noisy corpora from diverse sources.

Different Types of Events

If one plots the number of articles discussing an event against time, one gets characteristic shapes depending on the type of event:

- catastrophe
- expected event
- on-going event

Exercise: Check distribution of Usenet articles matching the query string "rolling blackout California" over time.

groups.google.com/advanced_group_search
find all terms: rolling blackout California
restrict to date range: e.g. 19 Mar 2001 to 19 Mar 2001

Result is
247 matches for 15 Jan 2001 through 14 Feb 2001, but only 20 hits for 15 Feb 2001 through 14 Mar 2001.

TDT Tasks: Segmentation

Given a stream of words, find the story boundaries.

Why?

- A news broadcast consists of multiple stories, but retrieving an entire half-hour show in response to a query is not useful.
- Different stories have different characteristics; for finding similarities, we don't want to be distracted by irrelevant stories in the same broadcast.
- News broadcasts contain non-news segments which we may wish to discard.

Approaches:

- decision trees
- language models
- Naive Bayes classifiers
- hidden markov models

TDT Segmentation: Approaches

Decision Trees

- Observation: certain key words and phrases (and other features) tend to occur almost exclusively near the beginning or end of a story, e.g.
  "<name> reporting from <location>"
- Training a decision tree to key off of such features will let us find the transition point from "end-of-story" features to "beginning-of-story" features and declare a boundary.

Language Models

- Observation: When a new story begins, it will use many terms which had not been used in the previous story.
- Training a static language model from a large number of news stories, we can compare its probabilities against those produced by a dynamic model updated as each word is processed. At a story boundary, there will be an abrupt shift in the ratio between the two models.

TDT Tasks: Online Detection

Assign a stream of news stories to appropriate categories.

Approaches:

- clustering
- multiple language models -- assign to topic with best-matching LM

TDT Tasks: Retrospective Detection

Group a collection of previously-received news stories into sets of similar documents.

Approaches:

- Clustering
    - single-pass
    - k-Means
    - agglomerative

TDT Tasks: Tracking

Given a small number of stories about an "interesting" event, find all future stories covering that event.

Approaches:

- as IR task: convert training stories into a query
- Decision Trees
- various document-similarity metrics
    - cosine similarity
    - k-Nearest Neighbors
- on-topic vs. background language models
- clustering

TDT Tasks: First-Story Detection

From a stream of news stories, pick out those with the very first mention of any event.

Approaches:

- basically use same systems as used for online detection

==> FSD can be done by ignoring all but the chronologically first story of the cluster for an event

TDT Tasks: Story Linking

Given two arbitrary news stories, determine whether or not they discuss the same event.

Approaches:

- term-vector similarity metrics
- clustering

Cross-Lingual TDT

Extra Challenging:

- stories are not just from different sources, they are in different languages as well!

Approaches:

- translation
  - full MT (e.g. Systran)
  - term substitution from bilingual dictionary
- multilingual processing -- if training data is available in multiple languages, treat each language independently
  - e.g. instead of clustering all languages together for detection, cluster each language separately and merge equivalent clusters afterwards

TDT Challenges

Some useful things that go beyond TDT:

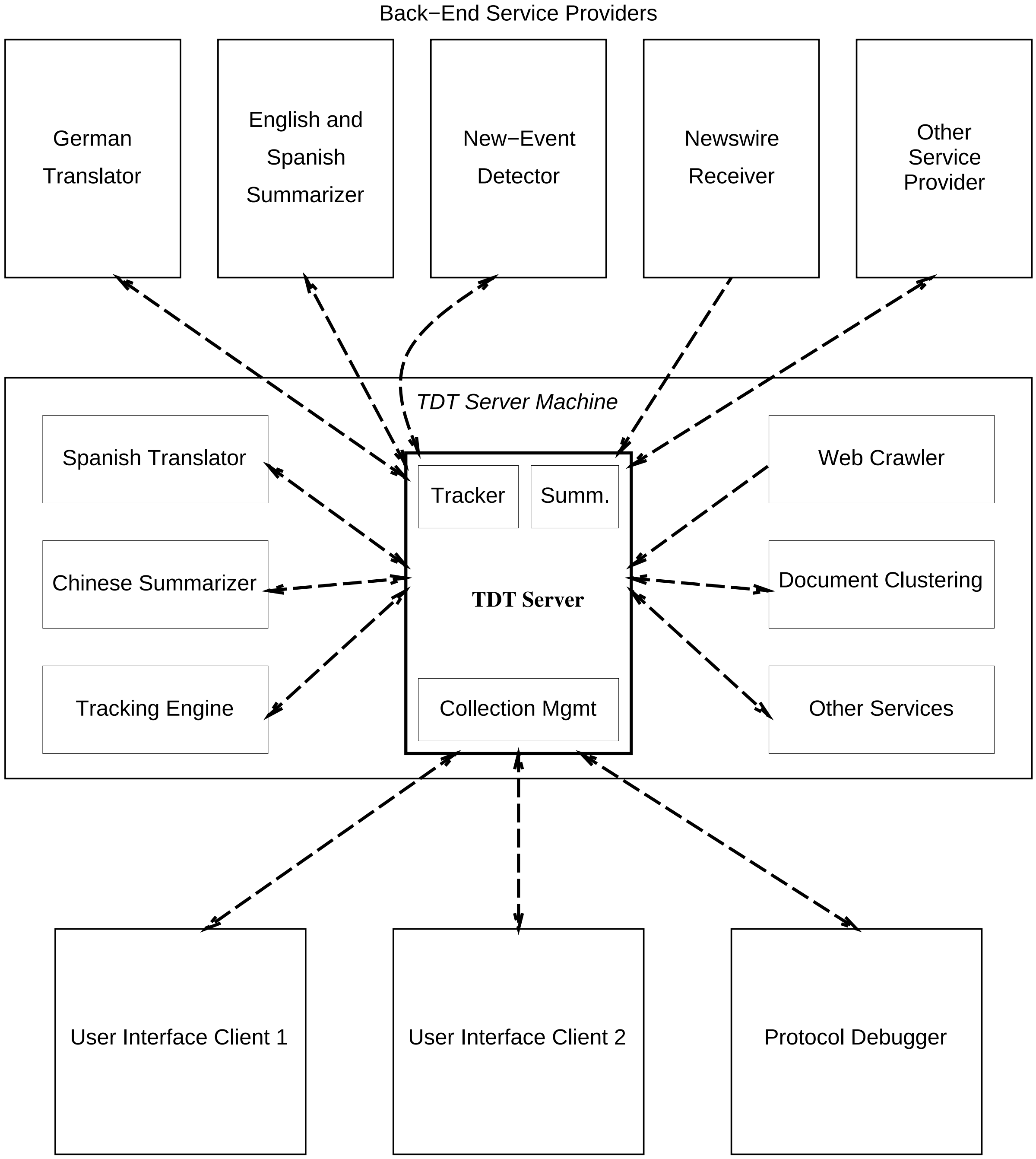
- Automatic Timeline Construction
  Problem: different sources often have conflicting info
- Tracking people/organizations across events
  Problem: lots of different ways to refer to the same entity:
  - Gerald Ford, President Ford, Ford
  - Ford Motor Company, Ford Motors, Ford
- Tracking relationships between people

Bringing TDT into the Real World

Now that the tasks are reasonably-well understood, the techniques need to be made available to end-users.

This means:

- real-time operation
- use of live newsfeeds
- simple-to-use user interfaces


Back−End Service Providers
German Translator
English and Spanish Summarizer
New−Event Detector
Newswire Receiver
Other Service Provider
TDT Server Machine
Spanish Translator
Chinese Summarizer
Tracking Engine
Tracker
Summ.
TDT Server
Collection Mgmt
Web Crawler
Document Clustering
Other Services
User Interface Client 1
User Interface Client 2
Protocol Debugger

Applying TDT to e-Commerce

- "clipping service" that can be provided to customers
- alerts when something happens that might affect your business