Automatic Timeline Generation from News Articles

Josh Taylor and Jessica Jenkins

Motivation

- Finding the major events in an ongoing story is difficult because news site searches will return results filled with only the events of the past two days.
- Example: a Google News search for "Iraq War" yields:
  - Rice's recent defense of the war
  - Recent polls showing low public support
- But it doesn't return results on:
  - Build-up to war
  - Major military operations
  - Lack of international support, U.N. controversy
  - Freedom fries
- Timeline presents major events in news story in an accessible format.

Language Model Approach

- Sentences from a set of articles on news story arranged chronologically.
- Construct a language model over sentences based on frequency counts and sentence ordering.
- Use model to score sentences for usefulness and novelty.
  - Usefulness: Sentence is on-topic for story, i.e., doesn't contain tangential information.
  - Novelty: Sentence presents information on a new event not covered by previous sentences.
- Highest scoring sentences are used for timeline.

Event-based Model

- Explicitly learn important events in a news story
- by clustering sentences.
- Select representatives from event clusters for timeline sentences.
- Explore various features for representing sentence vectors for clustering, including named entities, noun phrases, temporal cues.

Evaluation

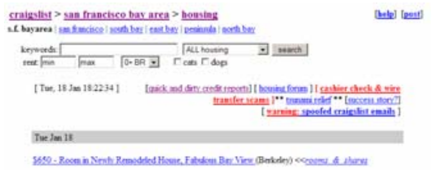
- Human annotators generate set of important events in news story.
- Each sentence is annotated with a (possibly empty) subset of the events it covers.
- Recall and precision measures based on these annotations are applied to the sequence of sentences returned by the system to evaluate the usefulness and novelty (or non-redundancy) of the timeline.

Information Extraction on Real Estate Rentals Classifieds

Eddy Hartanto
Ryohei Takahashi

Problem Definition

- craigslist.org is an online community
- Includes real estate postings
- But search is very basic:



craigslist > san francisco bay area > housing [help] [post]

s.f. bayarea | san francisco | south bay | east bay | peninsula | north bay

keywords: ALL housing search

rent: min max 0+ BR □ cats □ dogs

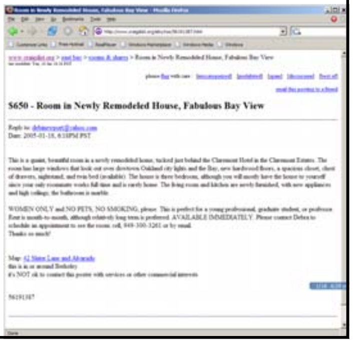
[Tue, 18 Jan 18:22:34] [quick and dirty credit reports] [housing forum] [cashier check & wire transfer scams] ** tsunami relief ** [success story?] [warning: spoofed craigslist emails]

Tue Jan 18

$650 - Room in Newly Remodeled House, Fabulous Bay View (Berkeley) <<rooms & shares

Problem

- Postings are unstructured
- Would be helpful to have structured information: e.g. deposit, refrigerator, square footage, etc.



www.craigslist.org > east bay > rooms & shares > Room in Newly Remodeled House, Fabulous Bay View

[illegible]

please flag with care: [miscategorized] [prohibited] [spam] [discussion] [best of]

email this posting to a friend

$650 - Room in Newly Remodeled House, Fabulous Bay View

Reply to: [illegible]

Date: 2005-01-18, 6:10PM PST

This is a quaint, beautiful room in a newly remodeled home, tucked just behind the Claremont Hotel in the Claremont Estates. The room has large windows that look out over downtown Oakland city lights and the Bay, new hardwood floors, a spacious closet, chest of drawers, nightstand, and twin bed (available). The house is three bedrooms, although you will mostly have the house to yourself since your only roommate works full time and is rarely home. The living room and kitchen are newly furnished, with new appliances and high ceilings; the bathroom is marble.

WOMEN ONLY and NO PETS, NO SMOKING, please. This is perfect for a young professional, graduate student, or professor. Rent is month-to-month, although relatively long term is preferred. AVAILABLE IMMEDIATELY. Please contact Debra to schedule an appointment to see the room: cell, 949-300-3261 or by email.
Thanks so much!

Map: 42 Slater Lane and Alvarado
this is in or around Berkeley
it's NOT ok to contact this poster with services or other commercial interests

56191387

Project Outline

- Crawl craigslist's real estate postings
- Extract structured information from unstructured text
- Offer parametric search on resulting database

Implementation Details

- Hidden Markov Model
  - States are fields
  - Outputs are words
  - Use Viterbi algorithm to calculate most likely sequence of states
- Rule-based pattern matching
  - Construct rules to identify words in postings that contain field data

Evaluation Measure

- Obtain random subset of postings
- Manually fill in fields of database for each of these postings
- Calculate precision/recall on a variety of queries on this set of manually tagged data

Questions and Suggestions

- We appreciate your inputs …

Web Crawling Stanford Events

Group members:

- Zoe Pi-Chun Chu
- Michael Tung

Scope

Building a school-wide events calendar.

Problem: information is separated, hard to maintain/update.

http://events.stanford.edu

-Requires manual input

-very few participating departments/student groups

Solution

An automated system

Builds events database by crawling:

-stanford.edu www pages

-newgroups

-mailing lists

Extract event attributes from text

(location, time, type, department, free food, speaker)

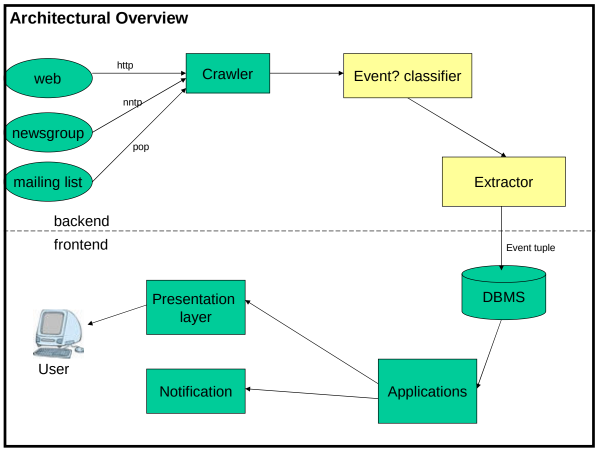
Technologies

Java Technology:
Build on Apache Tomcat

- JSP for dynamically generated webpages
- JavaBeans for data storage
- Java Mail API
- JDBC connects databases
- Lucene search engine

Databases:
MySQL

Architectural Overview



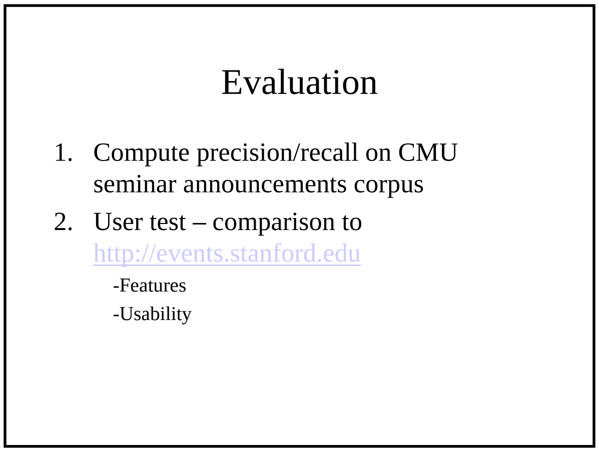
Key Algorithms

- Classification
  - For deciding whether content is an event
  - Segmenting events

Information Extraction

-Pattern matching, Part-of-speech tagging

-Hidden Markov model



Evaluation

1. Compute precision/recall on CMU seminar announcements corpus
2. User test – comparison to http://events.stanford.edu

   -Features

   -Usability