Community Events

Team C

Stanislaw Antol, Souleiman Ayoub, Carlos Folgar, Steve Smith

CS 4984 - Computational Linguistics
12/08/2014

Virginia Polytechnic Institute and State University
Blacksburg, VA

Outline

- Problem Statement
- Complications
- Approach
- Results
- Interesting Findings
- Lessons Learned
- Summary

Problem Statement

- Generate a summary of events within Blacksburg

This primarily requires being able to identify who, what, where, and when. This proved to be rather difficult.

Complications

- The number of event types are non-fixed and unknowable in their entirety ahead of time.
- Equivalent in some ways to handling multiple disasters of unknown type and quantity

Complications

- Time management issues
- Scheduling conflicts
- Differences in type and kind of data necessitating differences in approach - we were unable to apply the techniques most other groups were finding successful.

Complications

- Early errors lead to compounding issues later in the project
    - Proper data cleaning is a notable example - insufficiently cleaning the data early lead to nearly unusable results later on.

Approach

- Attempted a variety of filtering options
    - RegEx parsing
    - Classification
    - POS Tagging + CFG
    - Clustering
    - NER Tagging

Approach - Alternatives

- RegEx Parsing
  - Useful for templates, but we did not use templates
  - Used to help with NER tagging
- Classification
  - Previous work was unsatisfactory
  - Could be useful with more classifiers
- POS Tagging + CFG
  - Novel idea, but would take too long to implement

Approach - NER Tagging

- Determine if a sentence is “good”
  - Location
  - Date/Time - Using RegExp
  - Person/Organization
- Use with Clustering
  - Future refinement will make use of NER in conjunction with sub-clusters to find the best sentence based both on nearness to the centroid and existence of key named entity types.

Approach - Clustering

- Collection
    - Major Cluster (50 Total)
        - Sub-cluster (10 Per)
            - Sentence cluster (10 Per)
                - One sentence chosen from each sentence cluster
                - Closest to cluster centroid

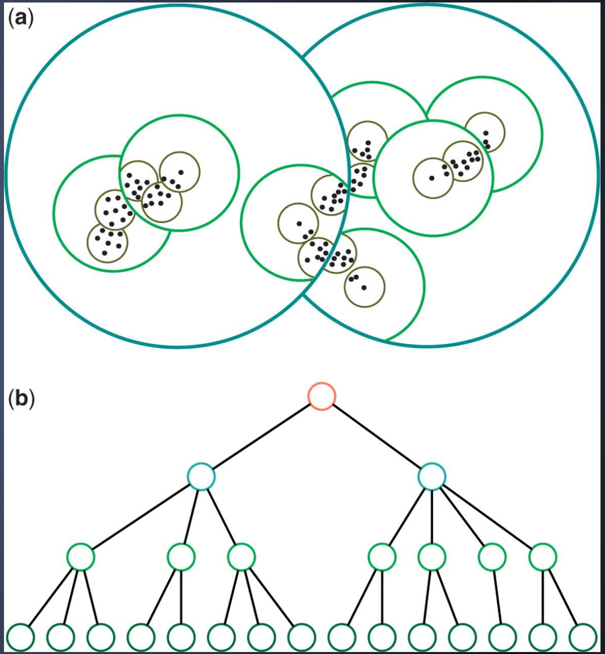

Image Credit: http://nar.oxfordjournals.org/content/early/2011/05/19/nar.gkr349/F1.large.jpg

Approach - Clustering

Settled on four major clusters based on observation.

- Local festivals, *etc*.
- Crime/Police
- Weather
- Virginia Tech events

Results

The following slides showcase some examples of the results we achieved by picking sentences close to the centroid of the deepest sub-clusters of each major cluster. With further refinement we feel confident we could extract a unique and distinct sentence that would summarize the event with clarity.

Results for Local/Festival Cluster

- Kick off the Spring Season and your Easter Holiday with fun craft activities as Main Street Radford hosts the annual Bunny Trail in East Downtown on Saturday March 31 2012 from 10:30 a.m. to 1:00 p.m.

- The Montgomery County Educational Foundation will presents its First Annual Benefit Concert at the Blacksburg Presbyterian Church on Saturday Feb 2 at 3pm.

Results for Crime/Police Cluster

- Police had investigated 16 reports of peeping and spying.Christiansburg police charged Walrond with three felonies -- breaking and entering grand larceny and the unlawful video recording of a child younger than 14 -- and one misdemeanor count of peeping into an occupied dwelling.

- A Pulaski County man will serve 15 years for his role in a drug-related hammer assault Radford Commonwealth Attorney Chris Rehak said in a release.Lorne Rivens 42 of Dublin was charged with armed burglary attempted robbery and malicious wounding stemming from an April 19 2011 incident according to the release.Rehak said Rivens went to Radford to settle a drug debt but got the address wrong.

Results for Weather Cluster

- Winter storm warnings above 3000 feet for Alleghany and Bath counties and west of Interstate 77 Cold air circulating around the storm as it moves inland combined with thick moisture wrapping around it will lead to heavy snow in higher elevations near the Virginia-West Virginia line.

- Power outages swept the area over the weekend leaving hundreds in Franklin Henry and Montgomery counties without power at times.

- Appalachian did not return calls Sunday to say when power may be restored or if the outages were related to high winds in the area.

Results for VT cluster

- Five-member team made up of faculty members from three different colleges has received the 2012 XCaliber Award for excellence as a group involved with technology-assisted teaching.

- Members of the Virginia Tech Rescue Squad placed first in the Advanced Life Support Skills competition at the National Collegiate Emergency Medical Services Foundation Conference in Baltimore on Feb 25-26.

Interesting Findings

- Used cosine distance measure
  - Originally used Euclidian distance ($c = \surd(a^2+b^2)$)
  - Found better results using cosine difference as a distance metric
- The sentences found were often one of the earlier sentences in a document
  - It makes sense to put important information about an event close to the beginning of documents about said event so that readers don’t have to search.

Lessons Learned / Summary

- Make sure data is clean
- Our topic was significantly different than other groups and therefore required a significantly different approach
- Multiple cluster levels yielded best results

Acknowledgments

Special thanks to:

- Dr. Fox for providing this opportunity
- Our TAs for their assistance throughout the semester
- The National Science Foundation for their grant (NSF DUE-1141209) which enabled this course to be offered