Automated Named Entity Extraction for Tracking Censorship of Current Events

Antonio M. Espinoza & Jedidiah R. Crandall
Computer Science Department
University of New Mexico

{amajest, crandall}@cs.unm.edu
Firewall???
Fire Department
What's Censored?

<table>
<thead>
<tr>
<th>Text</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>dajiyuan</td>
<td>Pinyin for Epoch Times</td>
</tr>
<tr>
<td>茉莉花革命</td>
<td>Jasmine Revolution</td>
</tr>
<tr>
<td>罢课</td>
<td>Strike</td>
</tr>
<tr>
<td>freedom</td>
<td>Freedom</td>
</tr>
<tr>
<td>请愿书</td>
<td>Petition</td>
</tr>
<tr>
<td>华夏论坛</td>
<td>China Forum</td>
</tr>
<tr>
<td>学潮</td>
<td>Student protests</td>
</tr>
</tbody>
</table>
April 22\textsuperscript{nd}, 1989
Stages

- Express grievances
- Build a social network
- Create organization/movement
- Plan protest
- Establish demands
- Advertisement
- Action
Stages

- Express grievances
- Build a social network
- Create organization/movement
- Plan protest
- Establish demands
- Advertisement
- Action
Censorship is dynamic
Tuesday, March 29th
Tuesday, June 28th
辐射
High-level design

- Download news daily
- Segment news
- Extract named entities
- Add named entities to list
- Probe using the list
Sources

- http://news.wenxuecity.com/
- http://www.epochtimes.com/
- http://www.cnd.org/
- http://www.popyard.org/
- http://www.rfa.org/

- Banned in China
High-level design

Download news daily

Segment news

Probe using the list

Add named entities to list

Extract named entities
上周获取保候审的北京艺术家艾未未本周一被税务局追罚巨款后，他的妻子路青星期二对本台介绍了丈夫遭软禁及近况，并称已委托两位律师代理此案。浦志强律师告诉记
者，已要求税务部门开听证会，解释处罚的理由和程序是否妥当，而现有材料与艾未未
High-level design

- Download news daily
  - Probe using the list
  - Add named entities to list
- Segment news
  - Extract named entities
President Obama flew to New York last night to attend three back-to-back fundraisers. The first event was held at the home of former New Jersey governor and Goldman Sachs chair, Jon Corzine. The New York Times described the $35,800-a-plate event as Obama’s first in a "kiss-and-make-up effort" with Wall Street donors. The Obama campaign is hoping to raise $1 billion for the 2012 election.
Named entity extraction

The opennlp.maxent package was originally built by Jason Baldridge, Tom Morton, and Gann Bierner
http://maxent.sourceforge.net/about.html
High-level design

- Download news daily
- Segment news
- Extract named entities
- Add named entities to list
- Probe using the list
The list

- Priority 0-15
- 0 means blacklisted somewhere
- Words start out at priority 1
- After each probe move up
- After 15 probes, fall off the list
- Search a priority every 12 hours
- Only GET request censorship affects the list
High-level design

- Download news daily
- Segment news
- Extract named entities
- Add named entities to list
- Probe using the list
1) GET request probing
1) GET request probing

The connection was reset

The connection to the server was reset while the page was loading.

- The site could be temporarily unavailable or too busy. Try again in a few moments.
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Firefox is permitted to access the Web.

Try Again
2) Search engine probing

http://www.baidu.com/
http://www.sogou.com/

GET /s?wd=free

dom HTTP/1.1

Host: www1.baidu.com
2) Search engine probing
Implementation of NEE
What is maximum entropy modeling

- Constraints
- Find the maximal entropy model that meets the constraints
Why use maximum entropy?

- Can use diverse features
  - Flexibility of multiple languages
- Assume as little as possible about things you have not seen before
  - This sets it apart from maximum likelihood estimation
Topological sort

For context: prev_1 prev_has_punctuation 2 is_dict_term Tverb verb has caps next 2 next_is_dict_term next_Iverb
NOT_NAME[1.0000] BEGINNING_NAME[0.0000] MIDDLE_NAME[0.0000] END_NAME[0.0000] COMPLETE_NAME[0.0000]

For context: prev_2 prev_is_dict_term prev_Tverb prev_verb prev_has_caps 2 is_dict_term Iverb next_7
NOT_NAME[0.9993] BEGINNING_NAME[0.0006] MIDDLE_NAME[0.0000] END_NAME[0.0000] COMPLETE_NAME[0.0000]

For context: prev_2 prev_is_dict_term prev_Iverb 7 next_1 next_has_punctuation
NOT_NAME[0.9993] BEGINNING_NAME[0.0000] MIDDLE_NAME[0.0000] END_NAME[0.0000] COMPLETE_NAME[0.0006]

For context: prev_7 1 has_punctuation next_1 next_is_dict_term
NOT_NAME[1.0000] BEGINNING_NAME[0.0000] MIDDLE_NAME[0.0000] END_NAME[0.0000] COMPLETE_NAME[0.0000]

For context: prev_1 prev_has_punctuation 1 is_dict_term next_2 next_is_dict_term next_noun
NOT_NAME[1.0000] BEGINNING_NAME[0.0000] MIDDLE_NAME[0.0000] END_NAME[0.0000] COMPLETE_NAME[0.0000]
Experimental methodology
Experimental methodology

Train

Test
## Maximum Entropy results

<table>
<thead>
<tr>
<th>Category</th>
<th>Specificity</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Places</strong></td>
<td>69.8%</td>
<td>96.3%</td>
<td>0.77%</td>
</tr>
<tr>
<td><strong>People</strong></td>
<td>83.4%</td>
<td>89.63%</td>
<td>0.42%</td>
</tr>
<tr>
<td><strong>Organizations</strong></td>
<td>88.4%</td>
<td>87.56%</td>
<td>0.28%</td>
</tr>
<tr>
<td>Text</td>
<td>Translation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dajiyuan</td>
<td>Pinyin for Epoch Times</td>
<td></td>
<td></td>
</tr>
<tr>
<td>罢课</td>
<td>Strike</td>
<td></td>
<td></td>
</tr>
<tr>
<td>freedom</td>
<td>Freedom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>请愿书</td>
<td>Petition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>华夏论坛</td>
<td>China Forum</td>
<td></td>
<td></td>
</tr>
<tr>
<td>学潮</td>
<td>Student protests</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Search engine sensitive words

<table>
<thead>
<tr>
<th>Text</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>茉莉花</td>
<td>Jasmine flower</td>
</tr>
<tr>
<td>诺贝尔</td>
<td>Nobel Prize</td>
</tr>
<tr>
<td>刘先生</td>
<td>Mr. Liu</td>
</tr>
<tr>
<td>挪威</td>
<td>Norway</td>
</tr>
<tr>
<td>七十七,77, 七七</td>
<td>77</td>
</tr>
<tr>
<td>王府井</td>
<td>Wangfujing</td>
</tr>
</tbody>
</table>

This is not a complete list
Conclusions

During the two months that we gathered data the GET request black list appeared to be static. Even the Jasmine Flower (茉莉花) keyword didn't appear as a blip on the GET request censorship.
Future Work

- Focus more on search engine and blogging censorship
- Improve the NEE
- Support other languages
- Add daily TOM-Skype list
Acknowledgments

- Anonymous FOCI reviewers
- Fletcher Hazlehurst, Veronika Strnadova, Leif Guillermo, Ronald Garduno, Terran Lane, Sergey Plis
- This material is based upon work supported by the National Science Foundation under Grant No. #0844880 and #1025447
Thank You

¿Questions?