Detecting Incidents with Limited Linguistic Knowledge for Low Resource Languages

EnvComp at ISGC 2017

Chao-Hong Liu, Qun Liu and Ching-Hsien Lee

ADAPT Centre, DCU, Ireland; NTU, Taiwan
Background

• ADAPT - The Centre for Digital Content Technology, Ireland
  – http://www.adaptcentre.ie/
  – Formerly known as CNGL
    (Centre for Next Generation Localisation)
  – Crisis MT (EU), kick-off in April, 2017

• Industrial Technology Research Institute (ITRI), Hsinchu, Taiwan

• Central Geological Survey, MOEA, Taiwan
  – 1999 Jiji earthquake
Related Projects

• NIST LoReHLT 2016
  – Low Resource human language technologies
• Crisis Machine Translation (Crisis MT)
• Exploitation of Social Media for Emergency Relief and Preparedness (SMERP)
  – microblogs posted during the earthquake in Italy in August 2016
  – Shared Tasks on text retrieval and text summarization
• NAPES (EU FP7)
  – Next Generation Analytical Platforms for Environmental Sensing
  – http://www.napes.eu/
  – Workshop 2: March 27-28, 2017 at Croke Park, Dublin
    • INSIGHT Centre
    • NCSR (National Centre for Sensor Research)
Use of Language Processing

In Sensing the Environments

• “Social media in Ebola outbreak,” Hossain et al.
  – doi: 10.1017/S095026881600039X

• “On the use of the Internet to collect earthquake information,” edited by Bossu and Earle

• Citizen science, “crowd” sensing

• etc.
An Open Participant

From: Chao-Hong Liu
Subject: Coordination on ADAPT Shared Tasks Participation

BRIEF3:
New (and Interesting) NIST evaluations for Low Resourced languages to register on May 16 . . .

9. NIST announced new Low resource human language technologies (LoReHLT) evaluation series
The LoReHLT 2016 Evaluations include three sessions:
1) Machine Translation (MT)
2) Named Entity Recognition (NER)
3) Situation Frame (SF)

Situation Frame is now being defined and can be briefed as to detect (the underlined values):
Household supply (Type) 99% for sure (TypeConfidence) in Dublin (EntityType=Location, Begin=2015, End=Ongoing) with Urgency (Status).

20-Mar-17
Team EarthSensory

• Chao-Hong Liu and Ching-Hsien Lee
• Professor: Qun Liu

• The DCU-NTU incident detection systems for LoReHLT16 SF task are designed as keyword spotting systems.
  – ChineseUyghurStarDict.txt (ZH to UY)
  – Google translate (EN to ZH to UY)
    • SetS: English Scenario Model
Task: Situation Frame (SF)

**Situation Frame** is now being defined and can be briefed as to detect (the underlined values):

Household supply (Type) 99% for sure (TypeConfidence) in Dublin (EntityType=Location, Begin=2015, End=Ongoing) with Urgency (Status).

---

20-Mar-17
Topic Modeling

Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented data that suggest the answer is not a simple number.

One researcher presented evidence that suggests the number of genes required to keep life going in a single-celled organism may be as few as 250, even less than the 755 genes in the yeast genome. The other researcher, a computational biologist, estimated that the human genome contains between 250,000 and 300,000 genes, a number that matches previous estimates.

Although the researchers’ methods are very different, their results are not all that far apart, especially in comparison to the 30,000 genes in the human genome, notes St. Andrews University’s John Reed.


https://twitter.com/elijah_meeks/status/458695317664194560
What we looked for?

3. Ranked with **TF-IDF** to choose keywords

2. Look at how these words existed in previous 3 weeks of **pre-incidents**

1. For each week in **post-incidents**, their content **words** are of interest

- We can’t find this “occurrence pattern” from pre/post incident posts/documents
Use of Native Informant (NI)

- Five hours consultation at the end of 2\textsuperscript{nd} week
  - Only CP3 used NI’s knowledge in the systems

- Asking NI for
  - Geographic indicator words (given en examples)
  - Keywords reviewing/editing (zh-uy, en-zh-uy)
  - Keyphrases translation (made-up en to uy)
  - Fundamental linguistic knowledge about Uyghur
<table>
<thead>
<tr>
<th>Geographic Indicators</th>
<th>Uyghur Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>city</td>
<td>شهر</td>
</tr>
<tr>
<td>county</td>
<td>ناحیه</td>
</tr>
<tr>
<td>residential yard</td>
<td>قورو-جاي</td>
</tr>
<tr>
<td>Autonomous County (Oblast)</td>
<td>ئوبلاست</td>
</tr>
<tr>
<td>province</td>
<td>ئولکه</td>
</tr>
<tr>
<td>county</td>
<td>دولت</td>
</tr>
<tr>
<td>(No corresponding English word)</td>
<td>تاغ</td>
</tr>
</tbody>
</table>

20-Mar-17
## NI: Keywords/Keyphrases

<table>
<thead>
<tr>
<th>NI edited</th>
<th>Keywords for Terrorism</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>تؤيوقسز هۇجوم</td>
<td>تؤيوقسز هۇجوم</td>
<td>袭击</td>
</tr>
<tr>
<td>(Attack)</td>
<td></td>
<td>(Attack)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English Phrases</th>
<th>Uyghur Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>medical resources are scarce</td>
<td>داۋەلەش ئۇسۇلكنلەرى يېتىشەتەسابەك</td>
</tr>
<tr>
<td>to implement the curfew</td>
<td>نەمزەرەھەنت قەلبەق</td>
</tr>
<tr>
<td>water supply is insufficient</td>
<td>سۇ تەمەنلەش يېتىشەتەسابەك</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>condition</th>
<th>check point</th>
<th>eqvl. class</th>
<th>min error rate</th>
<th>max error rate</th>
<th>num subms.</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>constrained</td>
<td>CP 3</td>
<td>SFType</td>
<td>1.147</td>
<td>6.105</td>
<td>5</td>
<td>2.298</td>
</tr>
<tr>
<td>Un-constrained</td>
<td>CP 3</td>
<td>SFType</td>
<td>1.505</td>
<td>5.197</td>
<td>2</td>
<td>5.197</td>
</tr>
</tbody>
</table>
First Thoughts

- Earthquakes
  - I saw these in situ
  - Prediction?
  - detection
- early warning
  - sensors
- Logistics
  - where
  - needs
  - emergency level - triage

https://www.youtube.com/watch?v=8cqQJCMcLWQ

http://www.webpages.uidaho.edu/~simkat/geol345_files/345lecture9.htm
Earthquakes

• Early warning system?
  – Japan, 2011
  – Taiwan, 2013
  • Hey, we have the PTT villagers’ earthquake early warning “system”

Japan has automatic system . . .
“Emergency Notification:——
Earthquake in Fukushima. Get prepared for strong shockwave.”

“Shockwave in 30 seconds!!!”

http://techsea.cc/2013/11/13/ptt-earthquake/
• “Is this right? — . . . the first thing you sensed earthquakes is NOT to run or hide, but TO POST IT on the internet?”

Oct 31, 2013, magnitude 6.2 quake, 38,282 posts in PTT.

Ranked 2nd popular in its Gossips board history.

This is why we have the PTT “human-powered” early warning “system.”
Earthquakes

- On the use of the Internet to collect earthquake information / ed. by Rémy Bossu and Paul S. Earle

  - U.S. Geological Survey

  "The 2014 earthquake in Napa was detected by USGS in **29 seconds** using Twitter data"

Earthquakes

• Beyond incident detection
  – logistics
  • News:
    “STOP sending aids . . .”
    Too much for some, and
    NONE for other regions.
• Social networks (SN):
  – Taichung needs water
  – Nantou needs 1,000 tents
  – Foods NOT needed in . . . the next week

http://atc.archives.gov.tw/921/c_01_01.swf
What to model?

• Leo Tolstoy
  – “All happy families are like one another; each unhappy family is unhappy in its own way.”

• Happy families?
  – Posting patterns?

• Unhappy families
  – Situation frames . . .
  – Annotations

https://en.wikipedia.org/wiki/Leo_Tolstoy
What to model/annotate?

- Each post/document?
- Each incident?


What to model?

- Phenomena specific to earthquakes?
  - Euro 2016 soccer tournament
  - Sentiment of all of the English tweets, which amounted to about **17 million in total**

[Image of chart showing tweet sentiment over time during Euro 2016, with specific events labeled]
Discussions

• Text itself is not the only source
  – date/time
  – who posted it? (who = an account in social media)
  – where does s/he live? (most co-occurred place)
  – how posts/documents (and, how these whos) are related/connected?

• How about, detecting
  – Situation of unknown types? No annotations . . .
  – Alien invasion?
LoReLEI Language Packs

- Haitian Creole
- Turkish
- Uzbek
- Tamil
- Hausa
- Kurmanji Kurdish
- Zulu

- Mandarin
- Lao
- Vietnamese
Thank you!

• Comments?