Social Media & Text Analysis
lecture 10 - Automatic Summarization for Twitter

CSE 5539-0010 Ohio State University
Instructor: Wei Xu
Website: socialmedia-class.org
I'm watching *true life* “I'm addicted to Internet” ... while I'm on mine lol

Okay these girls on *True Life I'm Too Beautiful* are not that pretty
Summarization

• Given a (or a set of) documents, generate a short summary

• Given a (large) set of topically and temporally clustered tweets, select a few representative tweets as the summary.
## Previous Work

<table>
<thead>
<tr>
<th>Selected Work</th>
<th>Size of Input</th>
<th>Length of Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wei et al. (2012)</td>
<td>average 10k tweets</td>
<td>10 tweets</td>
</tr>
<tr>
<td>Inouye &amp; Kalita (2011)</td>
<td>approximately 1500 tweets</td>
<td>4 tweets</td>
</tr>
<tr>
<td>Rosa et al. (2011)</td>
<td>average 410 tweets</td>
<td>1, 5, 10 tweets</td>
</tr>
<tr>
<td>Liu et al. (2011)</td>
<td>average 1.7k tweets</td>
<td>about 2 or 3 tweets</td>
</tr>
<tr>
<td>Takamura et al. (2011)</td>
<td>2.8k - 5.2k tweets</td>
<td>26 - 41 tweets</td>
</tr>
</tbody>
</table>

⚠️ Human annotators strongly prefer different numbers of tweets in a summary for different topics.

★ Used the length of human reference summaries to decide the length of system outputs, which information is not available in practice.
Research Questions

- What is the perfect length of multi-tweet summary?
- Will IE help summarization on Twitter?
  - noisy text: performance of IE?
  - short context: still need in-depth event analysis?
  - redundant: is word enough?
SumBasic

• Intuition:

words occurring frequently in the documents occur with higher probability in the human summaries than words occurring less frequently
SumBasic

- a very simple but strong summarization algorithm [Nenkova and Vanderwende, 2005]

- Intuition:

  words occurring frequently in the documents occur with higher probability in the human summaries than words occurring less frequently
SumBasic

• Step 1: computes the probability of each word $w$:

$$P(w) = \frac{n(w)}{\sum_i w_i}$$

• Step 2: computes the salience score of each sentence $S$:

$$Score(S) = \sum_{w \in S} \frac{P(w)}{|\{w \mid w \in S\}|}$$

• Step 3: pick the highest scored sentence into summary

• Step 4: for each word in sentences chosen at step 3, update their probability:

$$P_{new}(w) = P_{old}(w) \cdot P_{old}(w)$$

• Step 5: repeat Step 2~4 until reach desired length of summary
Varied-length Summary

• For a set of topically clustered tweets, amount of information varies greatly:
  • from very repetitive to very discrete
  • e.g.
  
  album release of a less notable singer
  vs.
  album release of a famous/controversy singer
Information Extraction (IE)

- Named Entity [Ritter et al. 2011 EMNLP]
- Event Phrases [Ritter et al. 2012 KDD]
Information Extraction (IE)

- **Named Entity** [Ritter et al. 2011 EMNLP]
- **Event Phrases** [Ritter et al. 2012 KDD]
Information Extraction (IE)

- **Named Entity** [Ritter et al. 2011 EMNLP]
- **Event Phrases** [Ritter et al. 2012 KDD]
- **Temporal Expressions** [Tabassum et al. 2016 EMNLP]

---

DRIVE News: Confused about the new road laws that came into effect on Jan1? This simplifies things #ygk

3:35 AM - 5 Jan 2016
Calendar Demo

- **Named Entity** [Ritter et al. 2011 EMNLP]
- **Event Phrases** [Ritter et al. 2012 KDD]
- **Temporal Expressions** [Tabassum et al. 2016 EMNLP]
- Count Entity/Day Co-occurrences (using $G^2$ Log Likelihood Ratio)
- Plot Top $k$ Entities for Each Day

http://statuscalendar.com
System Overflow

A lot of Tweets

Named Entity Recognition
Event Extraction

Clustered Tweets

Summary

Event Graph

Event Graph

Node - named entities + event phrase
Edge (weighted) - co-occurrence

Wei Xu, Alan Ritter, Ralph Grishman.
“A Preliminary Study of Tweet Summarization using Information Extraction” in LASM (2014)
Event Graph

Event Graph

Wei Xu, Alan Ritter, Ralph Grishman.

“A Preliminary Study of Tweet Summarization using Information Extraction” in LASM (2014)
PageRank

• a graph-based ranking algorithm
• a trademark of Google
• Idea: web surfing / random walk

The importance of a webpage is defined recursively and depends on the number and importance of all webpages that link to it.

• also used for local graph partitioning
PageRank

- Salience score of nodes:

$$ Score(u) = (1 - d) + d \times \sum_{v \in Adj(u)} \frac{Score(v)}{|Adj(v)|} $$

- directed graph
- iterate towards converge
- initial rank of node does not matter
- only edges matter
- total weight of the graph stays the same
PageRank → Event Rank

• Salience score of nodes:

\[
Score(u) = (1 - d) + d \times \sum_{v \in Adj(u)} \frac{e_{uv} \times Score(v)}{\sum_{w \in Adj(v)} e_{vw}}
\]

- undirected graph
- iterate towards converge
- initial rank of node does not matter
- only edges and their weights matter
- total weight of the graph stays the same
Graph Ranking

Graph Partitioning

- local graph partitioning by PageRank [Andersen et al., 2006]: a good partition of the graph can be obtained by separating high ranked vertices from low ranked vertices

Wei Xu, Alan Ritter, Ralph Grishman.
“A Preliminary Study of Tweet Summarization using Information Extraction” in LASM (2014)
Graph Partitioning

Graph Partitioning

Summary
Tweet1  Tweet2

Facebook
share

Instagram
account
delete

sell

Example Event Graph

Wei Xu, Alan Ritter, Ralph Grishman.
“A Preliminary Study of Tweet Summarization using Information Extraction” in LASM (2014)
Example Summary

| EventRank (Flexible)               | - So Instagram can sell your pictures to advertisers without you knowing starting January 16th I’m bout to delete my Instagram!
|                                 | - Instagram debuts new privacy policy, set to share user data with Facebook beginning January 16 |
| Instagram 1/16/2013               | - Instagram will have the rights to sell your photos to Advertisers as of Jan 16 |
| SumBasic                         | - Over for Instagram on January 16th |
|                                 | - Instagram says it now has the right to sell your photos unless you delete your account by January 16th http://t.co/tsjic6yA |

Example Event Graph

Figure 2: Event graph of 'Google - 1/16/2013', an example of event cluster with multiple focuses

**Example Summary**

<table>
<thead>
<tr>
<th>Google 1/16/2013</th>
<th>EventRank (Flexible)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Google’s home page is a Zamboni game in celebration of Frank Zamboni’s birthday January 16 #GameOn</td>
</tr>
<tr>
<td></td>
<td>Today social, Tomorrow Google! Facebook Has Publicly Redefined Itself As A Search Company <a href="http://t.co/dAevB2V0">http://t.co/dAevB2V0</a> via @sai</td>
</tr>
<tr>
<td></td>
<td>Orange says has it has forced Google to pay for traffic. The Head of the Orange said on Wednesday it had ... <a href="http://t.co/dOqAHhWi">http://t.co/dOqAHhWi</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SumBasic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tomorrow’s Google doodle is going to be a Zamboni! I may have to take a vacation day.</td>
</tr>
<tr>
<td></td>
<td>the game on google today reminds me of hockey #tooexcited #saturday</td>
</tr>
<tr>
<td></td>
<td>The fact that I was soooo involved in that google doodle game says something about this Wednesday #TGIW You should try it!</td>
</tr>
</tbody>
</table>
Research Questions

- What is the perfect length of multi-tweet summary?
  variable length

- Will IE help summarization on Twitter?
  - noisy text: performance of IE?
    summary is more readable and newsworthy
  - short context: still need in-depth event analysis?
    self-contained (no coref.) → better event graph
  - redundant: is word enough?
    unbalanced event graph → easier partitioning

Instructor: Wei Xu
www.cis.upenn.edu/~xwe/

Course Website: socialmedia-class.org