Social Media & Text Analysis
lecture 6 - Automatic Summarization for Twitter

Instructor: Wei Xu
Website: socialmedia-class.org
[Recap] NLP Pipeline

classification (Naïve Bayes)

Language Identification

Regular Expression

Tokenization

Part-of-Speech (POS) Tagging

Shallow Parsing (Chunking)

Named Entity Recognition (NER)

Stemming

Normalization

Sequential Tagging
Timeline of NLP on Microblogs

2010

- First Story Detection (Petrovic et al.)
- Geographic Variation (Eisenstein et al.)

2011

- Named Entity Recog. (Ritter et al.)
- POS Tagging (Gimpel et al.)
- Normalization (Han and Baldwin)
- Summarization (Liu et. al)

2012

- Open-Domain Event Extraction (Ritter et al.)
- Censorship Detection (Bamman et al.)
Timeline of NLP on Microblogs

- **Summarization** (Xu et al.)
- **Normalization** (Xu et al.)
- **Paraphrase Extraction** (Xu et al.)
- **POS** (Cherry and Guo)

- **Machine Translation** (Ling et. al.)
  - **POS** (Owoputi et al.)
- **Parsing Weibo** (Wang et. al.)
- **Parsing Twitter** (Kong et. al.)
- **Dialogue Modeling** (Sordoni et al.)
Twitter Usage / Company Facts

- **302M**: Monthly active users
- **500M**: Tweets sent per day
- **80%**: Active users on mobile
- **3,900**: Employees around the world
- **35+**: Offices around the world
- **77%**: Accounts outside the U.S.
- **33**: Languages supported
- **50%**: Employees are engineers

_All numbers approximate as of March 31, 2015._
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.
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 $S$:

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

• 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]

• Event Phrases [Ritter et al. 2012]
Information Extraction (IE)

- **Named Entity** [Ritter et al. 2011]
- **Event Phrases** [Ritter et al. 2012]
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?
A lot of Tweets

Named Entity Recognition
Event Extraction

Summary

Clustered Tweets

Event Graph

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

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:

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

- 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:

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

- undirected graph
- iterate towards converge
- initial rank of node does not matter
- only edges \textit{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
Graph Partitioning

Facebook

Instagram

share

account

delete

tsell
Graph Partitioning

Summary
Tweet1  Tweet2

Facebook
share

Instagram
account
dele
sell

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

Figure 3: Event graph of 'Instagram - 1/16/2013', an example of event cluster with a single but complex focus

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

<table>
<thead>
<tr>
<th>EventRank (Flexible)</th>
<th>Instagram 1/16/2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>- So Instagram can sell your pictures to advertisers without u knowing starting January 16th I’m bout to delete my instagram!</td>
<td></td>
</tr>
<tr>
<td>- Instagram debuts new privacy policy, set to share user data with Facebook beginning January 16</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SumBasic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Instagram will have the rights to sell your photos to Advertisers as of Jan 16</td>
<td></td>
</tr>
<tr>
<td>- Over for Instagram on January 16th</td>
<td></td>
</tr>
<tr>
<td>- Instagram says it now has the right to sell your photos unless you delete your account by January 16th <a href="http://t.co/tsjic6yA">http://t.co/tsjic6yA</a></td>
<td></td>
</tr>
</tbody>
</table>
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>EventRank (Flexible)</th>
<th>SumBasic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google 1/16/2013</td>
<td></td>
</tr>
<tr>
<td>- Google ’s home page is a Zamboni game in celebration of Frank Zamboni ’s birthday January 16 #GameOn</td>
<td></td>
</tr>
<tr>
<td>- Today social, Tomorrow Google! Facebook Has Publicly Redefined Itself As A Search Company <a href="http://t.co/dAevB2V0">link</a> via @sai</td>
<td></td>
</tr>
<tr>
<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/dOqAHhW">link</a></td>
<td></td>
</tr>
<tr>
<td>- Tomorrow’s Google doodle is going to be a Zamboni! I may have to take a vacation day.</td>
<td></td>
</tr>
<tr>
<td>- the game on google today reminds me of hockey #tooexcited #saturday</td>
<td></td>
</tr>
<tr>
<td>- The fact that I was soooo involved in that google doodle game says something about this Wednesday #TGIW You should try it!</td>
<td></td>
</tr>
</tbody>
</table>

[Recap] 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?
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
Thank You!

Instructor: Wei Xu

www.cis.upenn.edu/~xwe/

Course Website: socialmedia-class.org