MediaMeter: A Global Monitor for Online News Coverage

Tadashi Nomoto
National Institute of Japanese Literature
Finding novel topics in news streams

So far not much success in the literature (extraction, machine learning)
Frequencies of (manually assigned) topic descriptors that appeared in the New York Times from June to December, 2013.
|    | ≤ 2 | ≤ 5 | ≤ 10 |
|----|-----|-----|------|
| 1  | 42.3% | 62.0% | 78.9% | 87.4% |
| Frequency > 10 : 12.6% |
SVM cannot handle a huge taxonomy (Liu, 2005)

The number of unique topics in NYT over 6 months exceeds 8,000.
Approach

Memory Based Topic Label Generation

WikiLabel
How it works: Overview

1. Look up Wikipedia to find pages most relevant to a news story

2. Generate label candidates from page titles

3. Pick those that are deemed most fit to represent the content
WikiLabel: Concept Generation with Wikipedia
WikiLabel: Concept Generation with Wikipedia
WikiLabel: Concept Generation with Wikipedia
Mechanics

\[ l^* = \arg \max_{l:p[l] \in \mathcal{U}} Prox(p[l], \hat{\theta}|_N), \]

where:

\[ Prox(p[l], \hat{\theta}|_N) = \lambda Sr(p[l], \hat{\theta}|_N) + (1 - \lambda) Lo(l, \hat{\theta}) \]

- \( Sr(p[l], \hat{\theta}|_N) \): content similarity
- \( Lo(l, \hat{\theta}) \): relevance of label
\[ S_{r}(r, q) = \left(1 + \sum_{t}^{N} (q(t) - r(t))^2 \right)^{-1} \]

\[ L_{o}(l, v) = \frac{\sum_{i}^{|l|} I(l[i], v)}{|l|} - 1 \]

\[ I(w, v) = \begin{cases} 1 & \text{if } w \in v \\ 0 & \text{otherwise.} \end{cases} \]
What if Wikipedia does not know the event ....

Use sentence compression to generalize
| 2009 detention of American hikers by Iran |
|------------------------------------------|
| detention                               |
| detention by Iran                       |
| detention of hikers                     |
| detention of hikers by Iran             |
| detention of American hikers by Iran    |
| 2009 detention                          |
| 2009 detention by Iran                  |
| 2009 detention of hikers                |
| 2009 detention of hikers by Iran        |
| 2009 detention of American hikers by Iran |

Making it shorter makes it more general
2009 detention of hikers by Iran of American
Use every NP in the title as a resource
2009 detention of American hikers by Iran

What you get with extension

Original approach

you start here
Testing it out in the field

| COUNTRY       | MEDIA OUTLET(S)                                      | #outlets | #stories       |
|---------------|------------------------------------------------------|----------|----------------|
| US/UK         | THE NEW YORK TIMES, YAHOO, CNN, MSNBC, FOX, WASHINGTON POST, ABC, BBC, REUTERS | 9        | 2,230 (239,844) |
| SOUTH-KOREA   | JOONGANG ILBO (English edition), CHOSUN ILBO (English edition) | 2        | 2,271 (19,008)  |
| JAPAN         | ASAHI, JCAST, JJI.COM, MAINICHI, NHK, NIKKEI, SANKEI, TBS, TOKYO, TV-ASAHI, YOMIURI | 11       | 2,815 (259,364) |
North-Korean Agenda

Topic Popularity (South Korea)

North-Korean nuclear issue: 0.585
Workers' Party of Korea: 0.073
North-Korean defectors: 0.071
North-Korean abductions of Japanese citizens: 0.049
Yeonpyeongdo>>bombardment: 0.024
North-Korean missile test: 0.023
North-Korea United-States relations: 0.022
North-Korean nuclear test: 0.015
Human rights in North-Korea: 0.014
North-Korean abductions: 0.013
North-Korea sponsored schools in Japan: 0.01
First Secretary of the Workers' Party of Korea: 0.008
Prisons in North-Korea: 0.005
North-South Summit: 0.005
North Korean abductions of Japanese: 0.003
Mount Kumgang>>Tourist Region: 0.003
Korean Language: 0.003
North-Korean Intelligence Agencies: 0.002

Topic Popularity (Japan)

Abductions of Japanese: 0.427
North-Korean nuclear issues: 0.31
Ryongchon disaster: 0.317
Kim Jong-il's visit to China: 0.079
Culture in North-Korea: 0.082
North-South relations: 0.21

Topic Popularity (US)

North-Korea relations: 0.144
North-Korea nuclear program: 0.09
North-Korean defectors: 0.069
North-Korea Russia relations: 0.035
North-Korea South-Korea relations: 0.024
North-Korean weapons of mass destruction: 0.017
North-Korea United-States relations: 0.016
North-Korea program: 0.015
North-Korean missile test: 0.014
North-Korean test: 0.011
North-Korean floods: 0.008
North-Korean nuclear test: 0.006
North-Korean famine: 0.003
rocket North-Korea: 0.001
province North-Korea: 0.001
North-Koreans: 0.001
North-Korean defectors: 0.001
North-Korean nuclear test: 0.001
North-Korea South-Korea relations: 0.001
North-Korean defectors: 0.001
North-Korean nuclear program: 0.001
North-Korea relations: 0.001
North-Korea Russia relations: 0.001
North-Korea nuclear issue: 0.001

News Coverage Ratio
Title is one of major topics in Article. Article gives a particular attention to Title.

Part of Article deals with Title. Article makes a clear reference to Title.

Part of Title has some relevance to a dominant theme of Article. Example: Title ‘European Tax System’ is partially relevant to an article discussing US Tax System.

Article makes a reference to part of Title.

Title has no relevance to Article, in whatever way.

| LANGUAGE    | RATING | #instances |
|-------------|--------|------------|
| ENGLISH     | 4.63   | 97         |
| JAPANESE    | 4.41   | 92         |
Trending Topics in the US Online News Media

- Iraqi insurgency
- Argentine debt restructuring
- The FIFA World Cup
- Malaysia
- Ukraine
- Islamic State of Iraq and the Levant
- The Syrian Civil War
- Bowe Bergdahl
- Crisis
- Pakistan

Dates: Jul 10 to Jul 23
Trending Topics in the US Online News Media

The FIFA World Cup
Evaluation Metric: ROUGE-W

| $s_1$                                | $s_2$                                | ROUGE-W |
|--------------------------------------|--------------------------------------|---------|
| The United States of America         | The United States of America         | 1       |
| The United States States             | The United States of America         | 0.529   |
| States                               | The United States of America         | 0.077   |

$$S(C|_k, l) = \frac{1}{k} \sum_{c \in C|_k} \text{ROUGE-W}(c, l)$$
Results

Text Rank vs. WikiLabel

|        | TRANK | RM₀   | RM₁   | RM₁/X |
|--------|-------|-------|-------|-------|
| NYT    | 0.000 | 0.056 | 0.056 | 0.069 |
| TDT    | 0.030 | 0.042 | 0.048 | 0.051 |
| FOX*   | 0.231 | 0.264 | 0.264 | 0.298 |

New York Times (2013) : 19,952
TDT (1994) : 15,863
FOX (2015) : 11,014
Wikipedia (2012)
Summary

- Talked about topic detection using WikiLabel
- Leveraging Wikipedia
- Generalizing concept with sentence compression
- Use of sentence compression led to a huge improvement, producing performance twice as good as that of TextRank
- Online topic learning seems promising
Frequencies of (manually assigned) topic descriptors that appeared in the New York Times from June to December, 2013.