MediaMeter: A Global Monitor for Online News Coverage

Tadashi Nomoto
National Institute of Japanese Literature

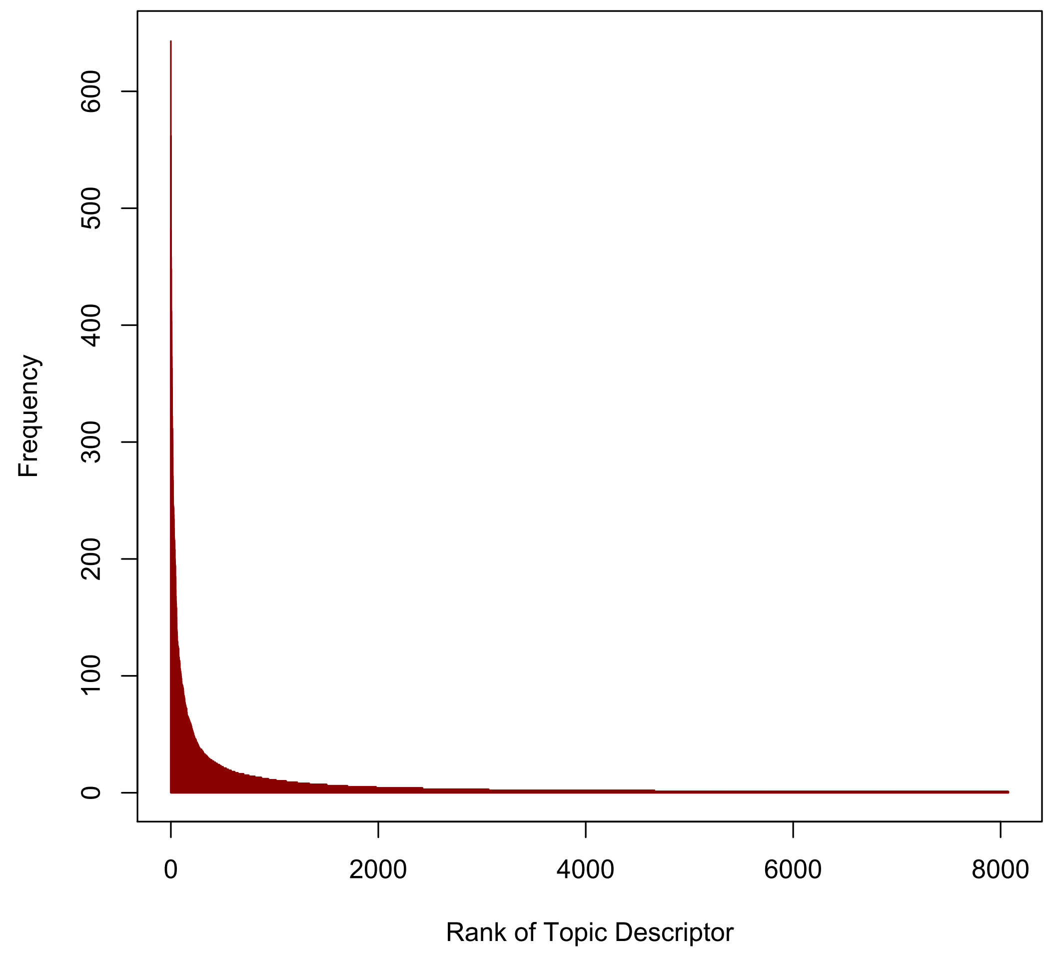
What we are aiming at

Finding novel topics in news streams

So far not much success in the literature
(extraction, machine learning)

Problem

Frequencies of (manually assigned) topic descriptors that appeared in the New York Times from June to December, 2013.



1	≤ 2	≤ 5	≤ 10
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.

Memory Based Topic Label Generation

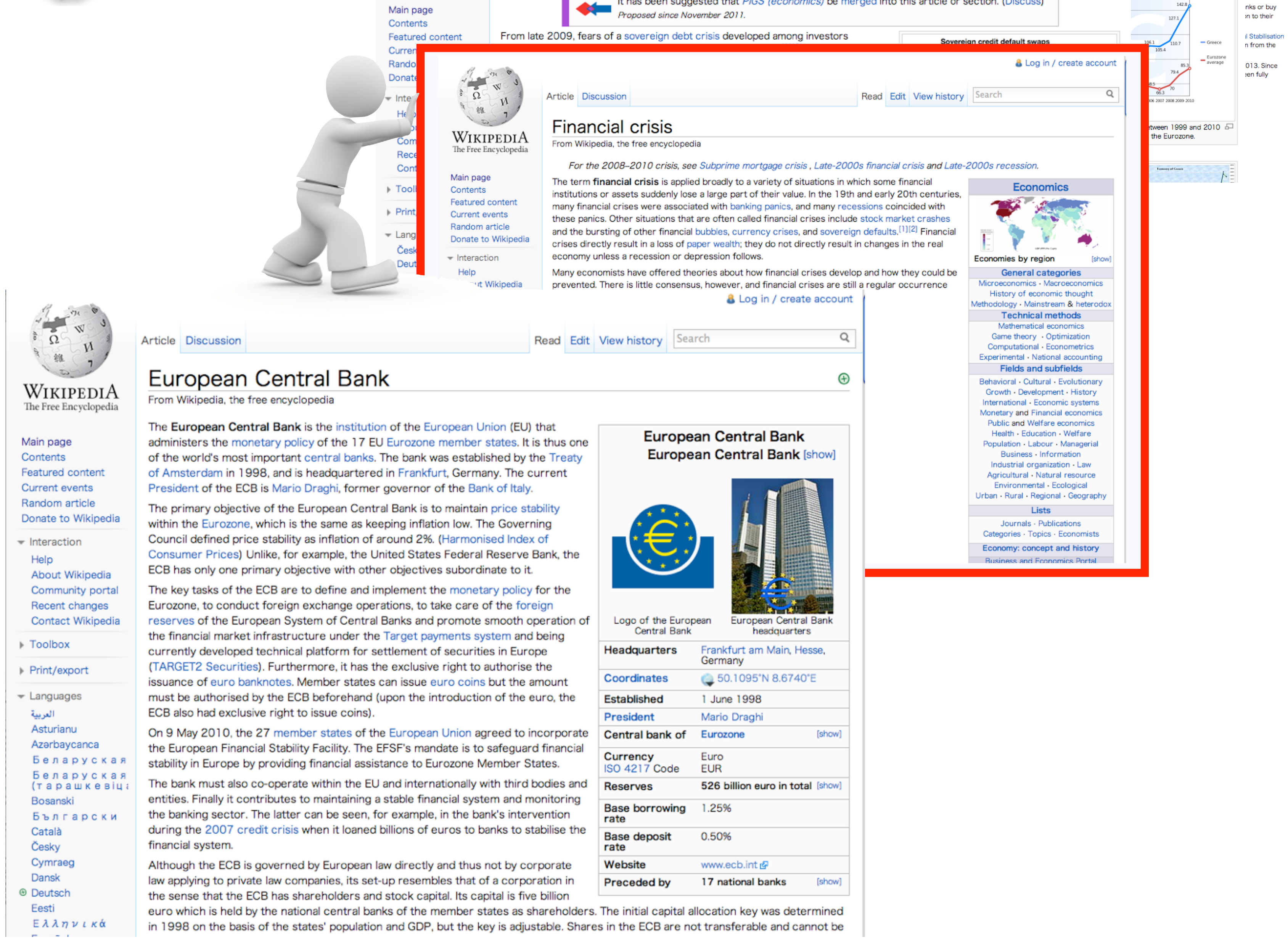
WikiLabel

How it works: Overview

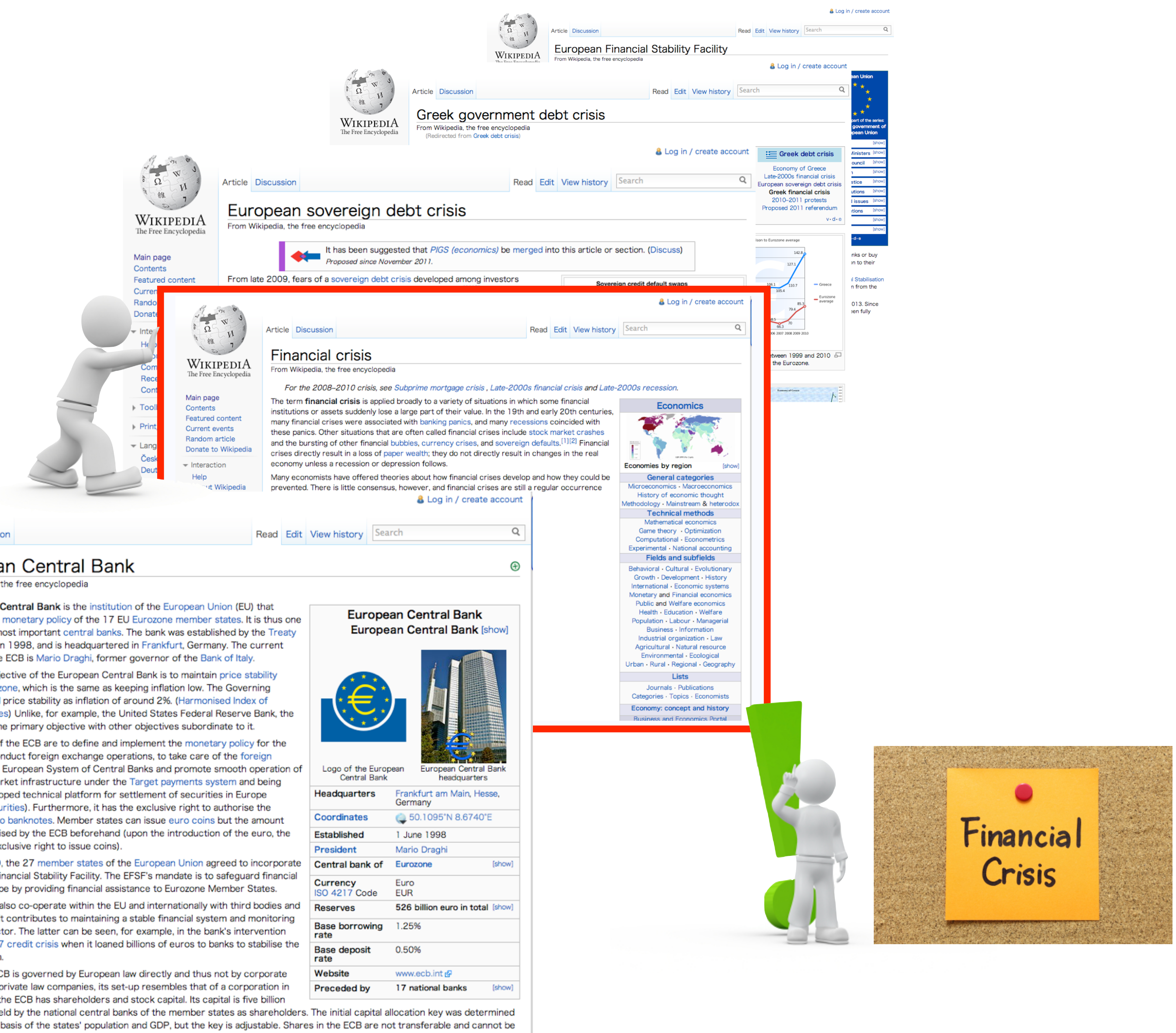
1. Look up Wikipedia to find pages most relevant to a news story
2. Generate label candidates from page titles
3. Pick those that are deemed most fit to represent the content


picture
behind
those
spring
whisper
time
leave
moment

WikiLabel: Concept Generation with Wikipedia



WikiLabel: Concept Generation with Wikipedia



Mechanics

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

news story *Concept Dictionary*

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

content similarity *relevance of label*

$$Sr(\mathbf{r},\mathbf{q}) = \left(1 + \sum_{t}^{N} (\mathbf{q}(t) - \mathbf{r}(t))^2\right)^{-1}$$

$$Lo(l,\mathbf{v}) = \frac{\sum_{i}^{|l|} I(l[i],\mathbf{v})}{\mid l \mid} - 1$$

$$I(w,v) = \begin{cases} 1 & \text{if } w \in v \\ 0 & \text{otherwise.} \end{cases}$$

What if Wikipedia does not know the event ....

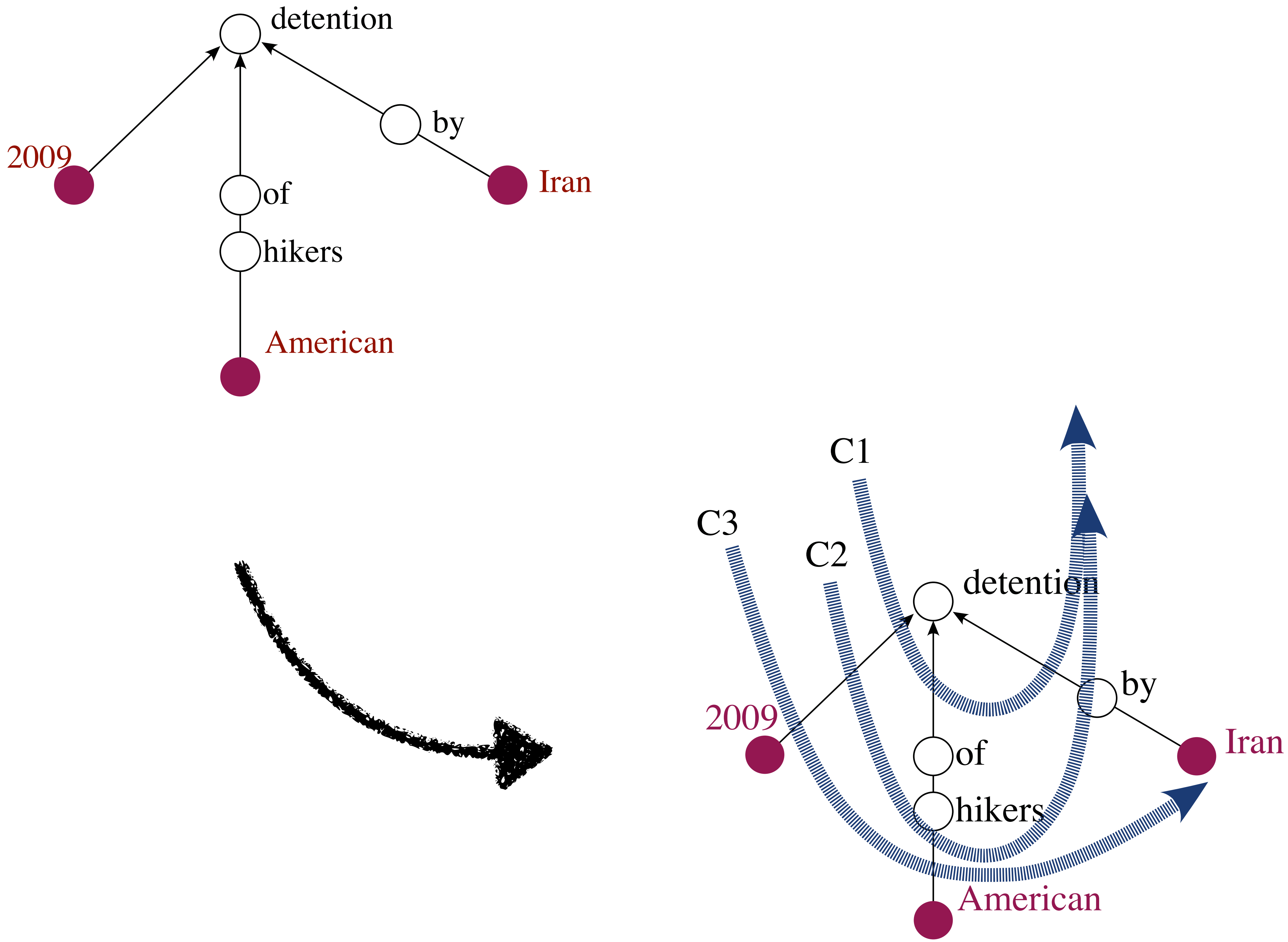
Use sentence compression to generalize

Example

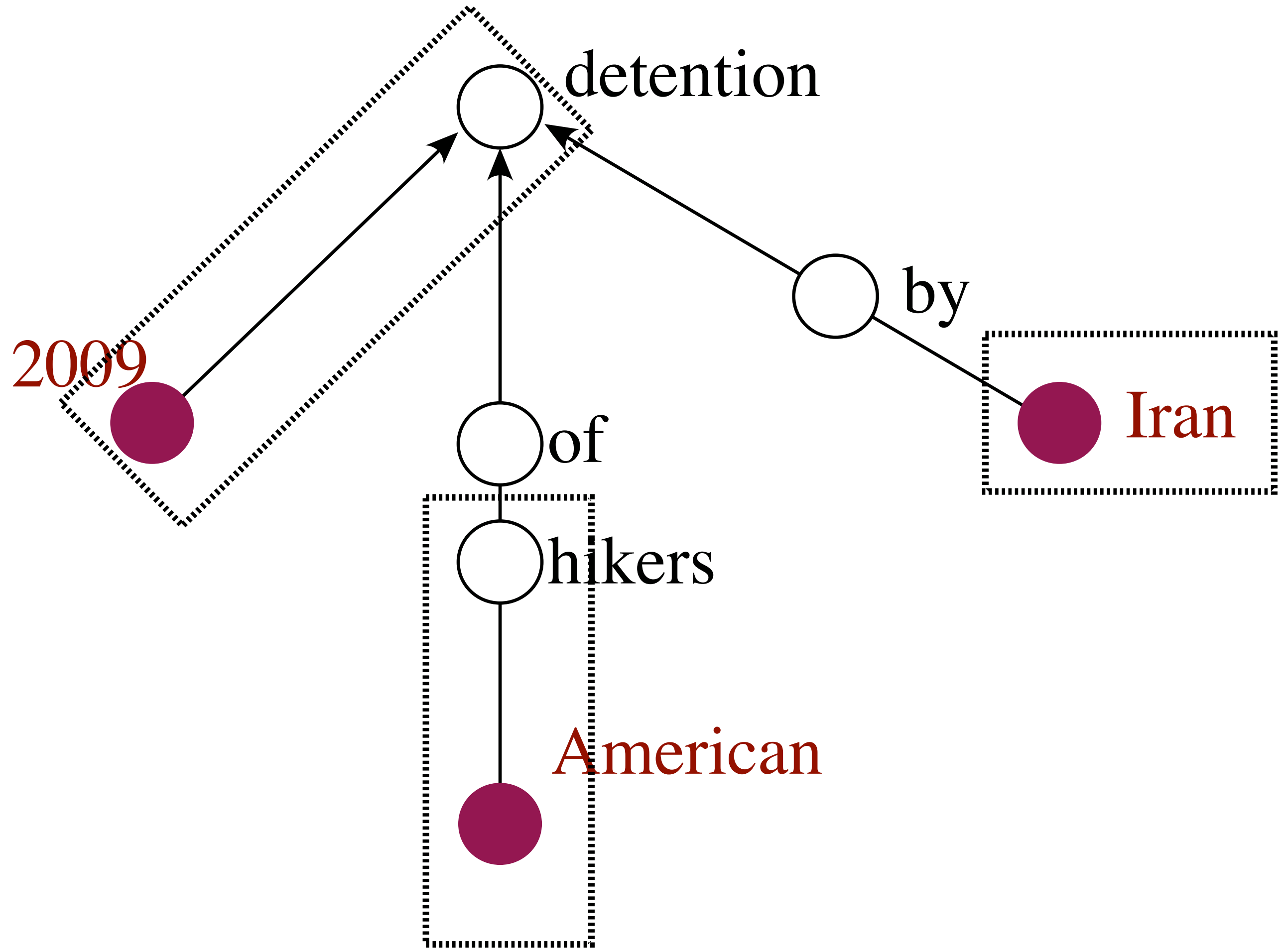
2009 detention of American hikers by Iran
detention
detention by Iran
detention of hikers
detention of hikers by Iran
detention of American hikers by Iran
2009 detention
2009 detention by Iran
2009 detention of hikers
2009 detention of hikers by Iran
2009 detention of American hikers by Iran

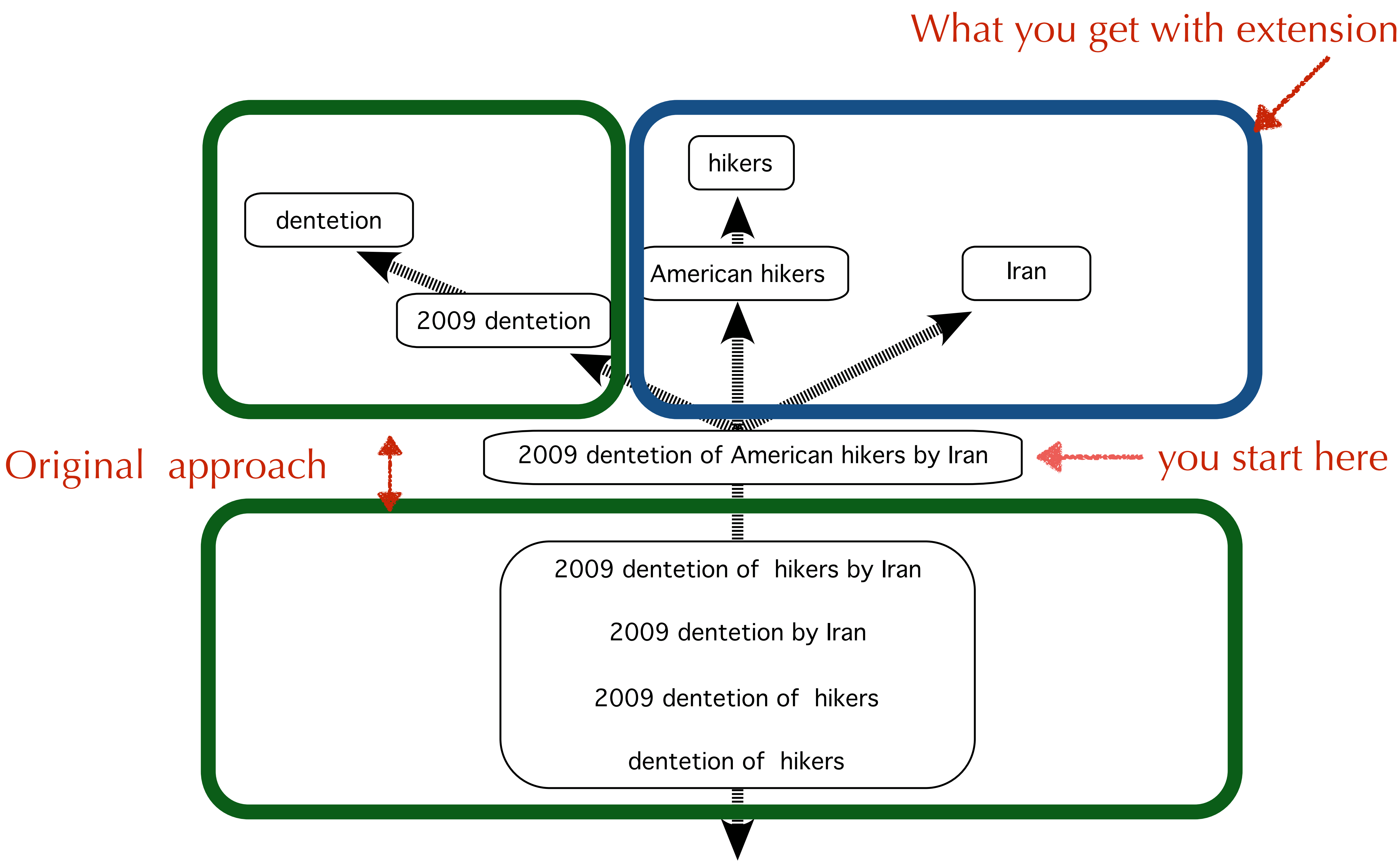
Making it shorter makes it more general

Dependency pruning



Use every NP in the title as a resource

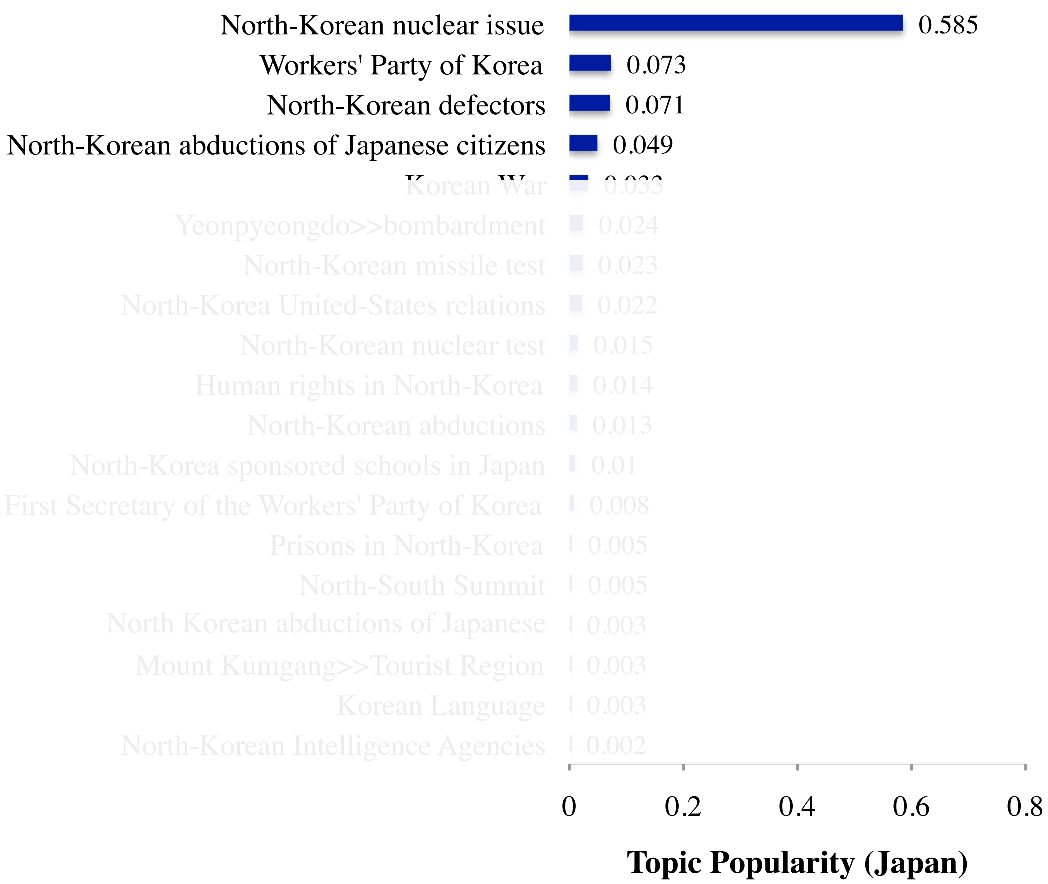


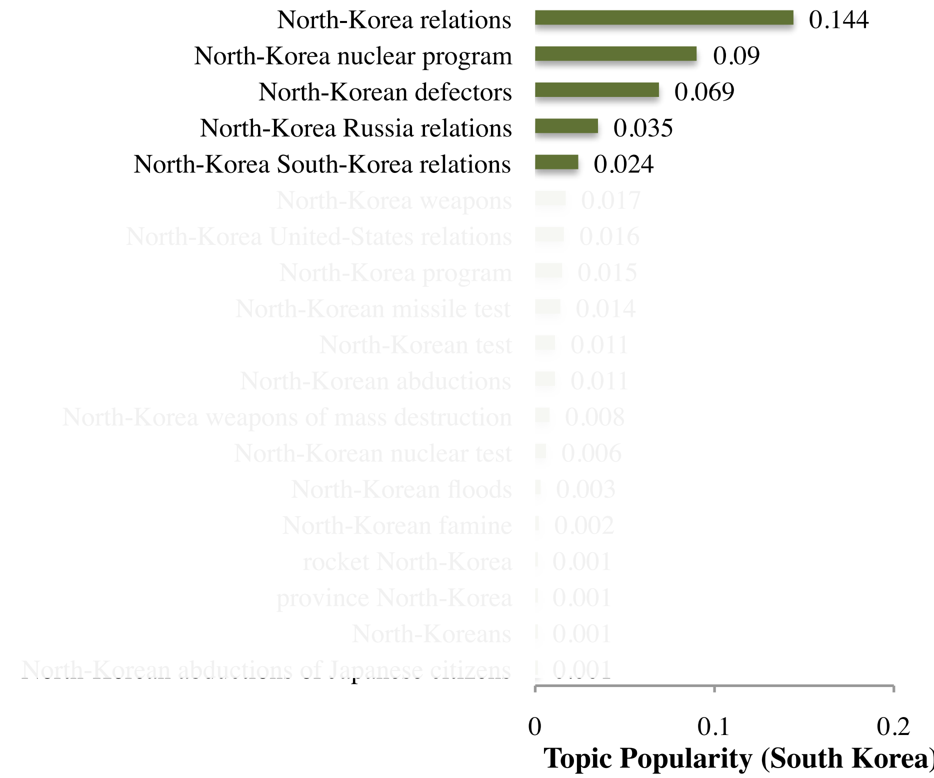

What you get with extension
dentetion
2009 dentetion
hikers
American hikers
Iran
Original approach
2009 dentetion of American hikers by Iran
you start here
2009 dentetion of hikers by Iran
2009 dentetion by Iran
2009 dentetion of hikers
dentetion of hikers

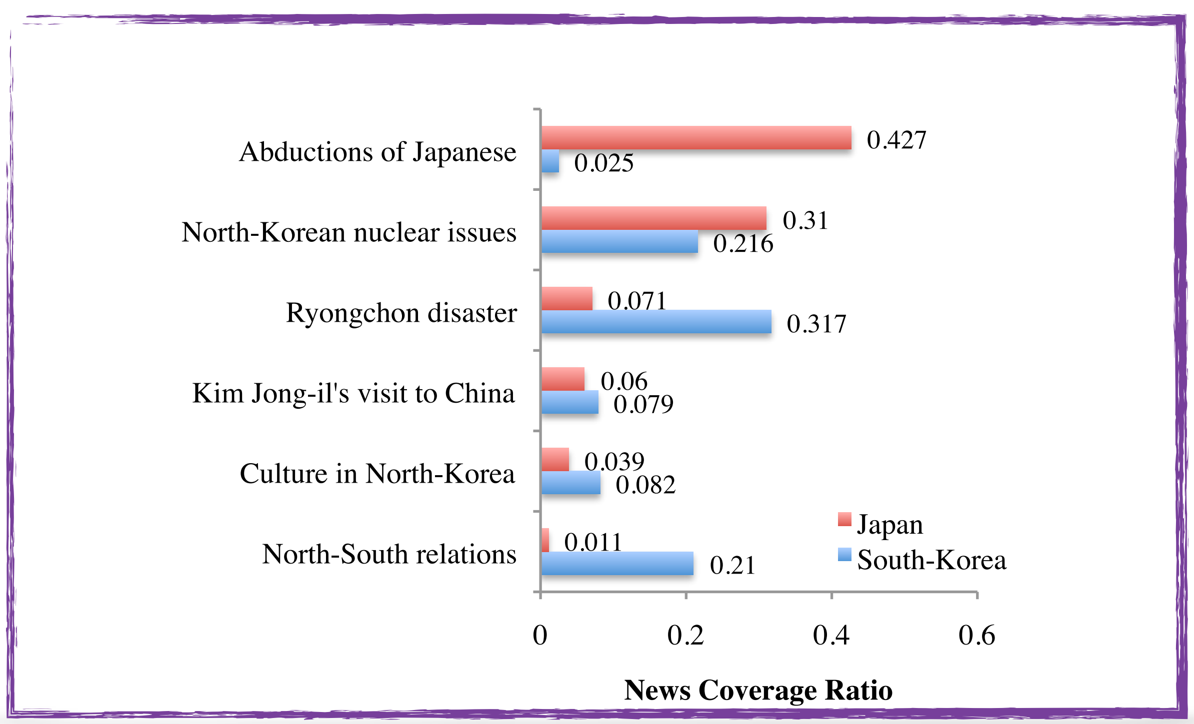
Testing it out in the field

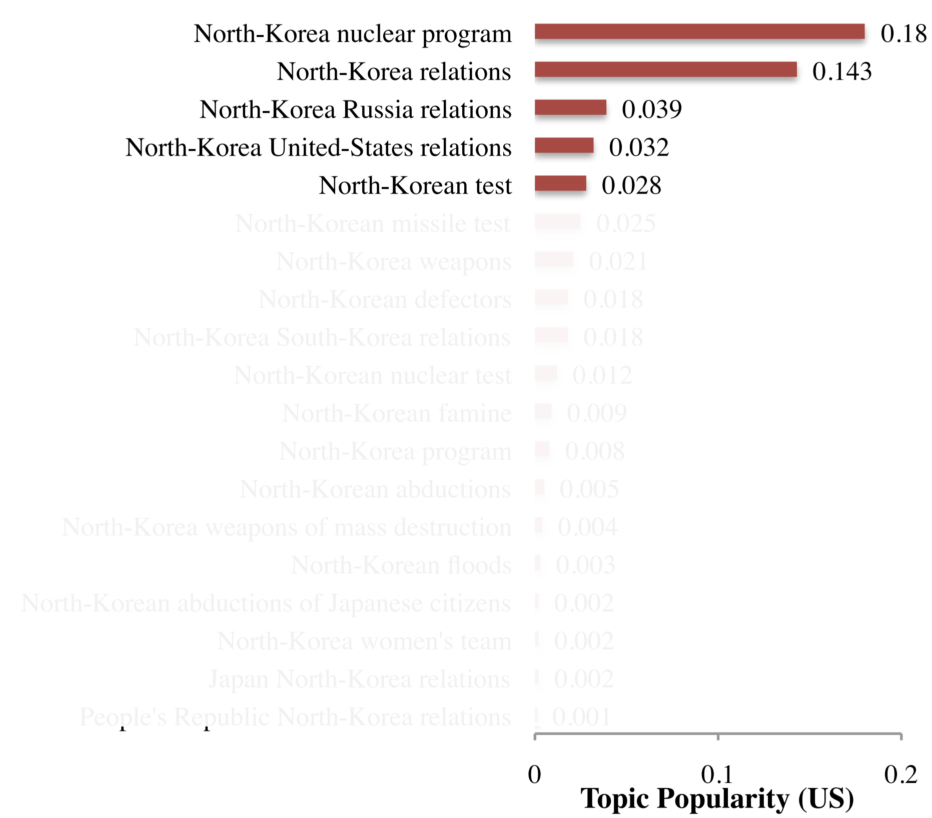
COUNTRY	MEDIA OUTLETS	#outlets	#stories
US/UK	THE NEW YORK TIMES, YAHOO, CNN, MSNBC, FOX, WASHINGTON POST, ABC, BBC, REUTERS	9	2,230 (239,844)
SOUTH-KOREA	JOONGANG ILBO (English edition), CHOSUN ILBO (English edition)	2	2,271(19,008)
JAPAN	ASAHI, JCAST, JIJI.COM, MAINICHI, NHK, NIKKEI, SANKEI, TBS, TOKYO,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

North-Korean Agenda







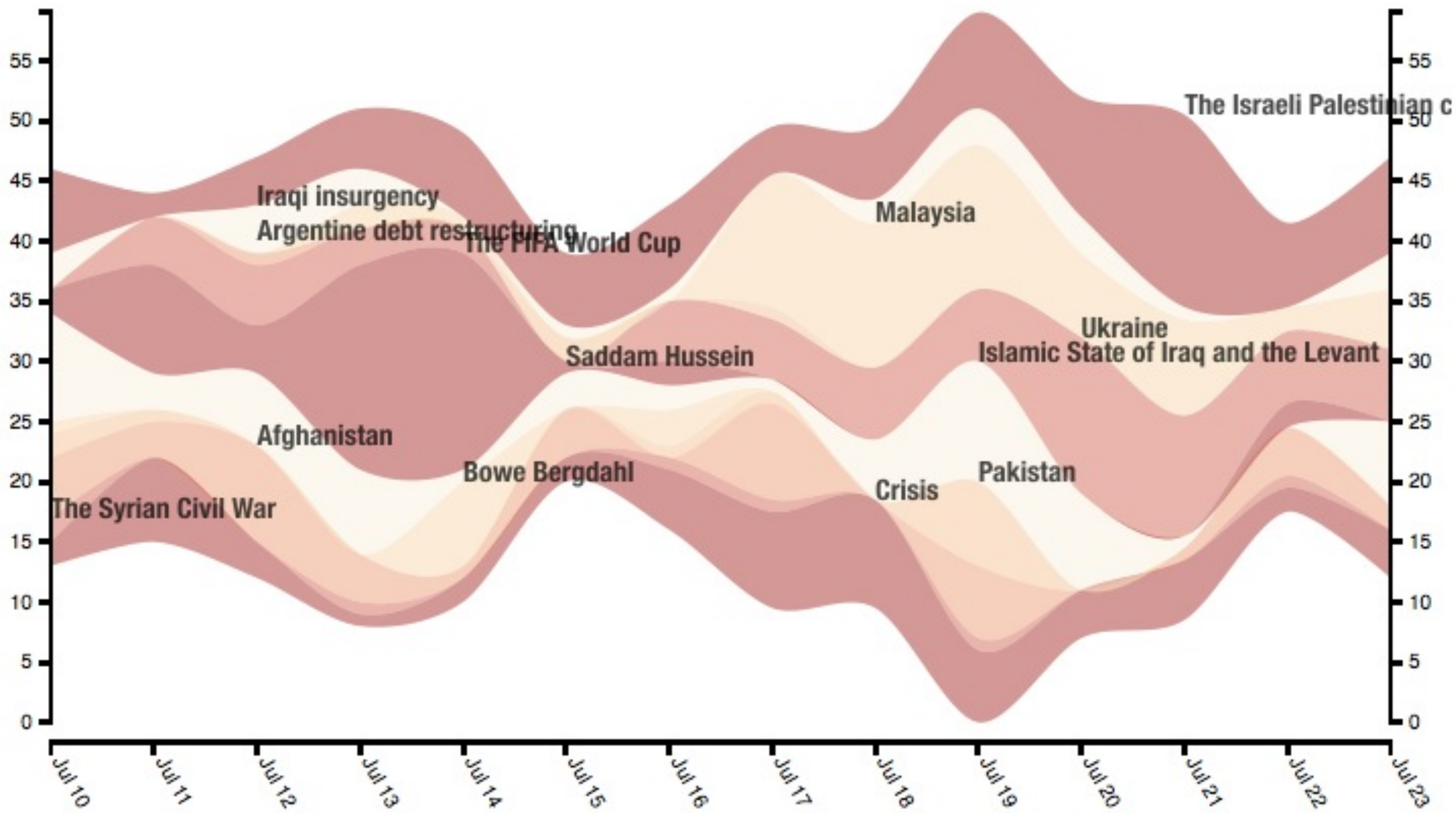


Human Evaluation

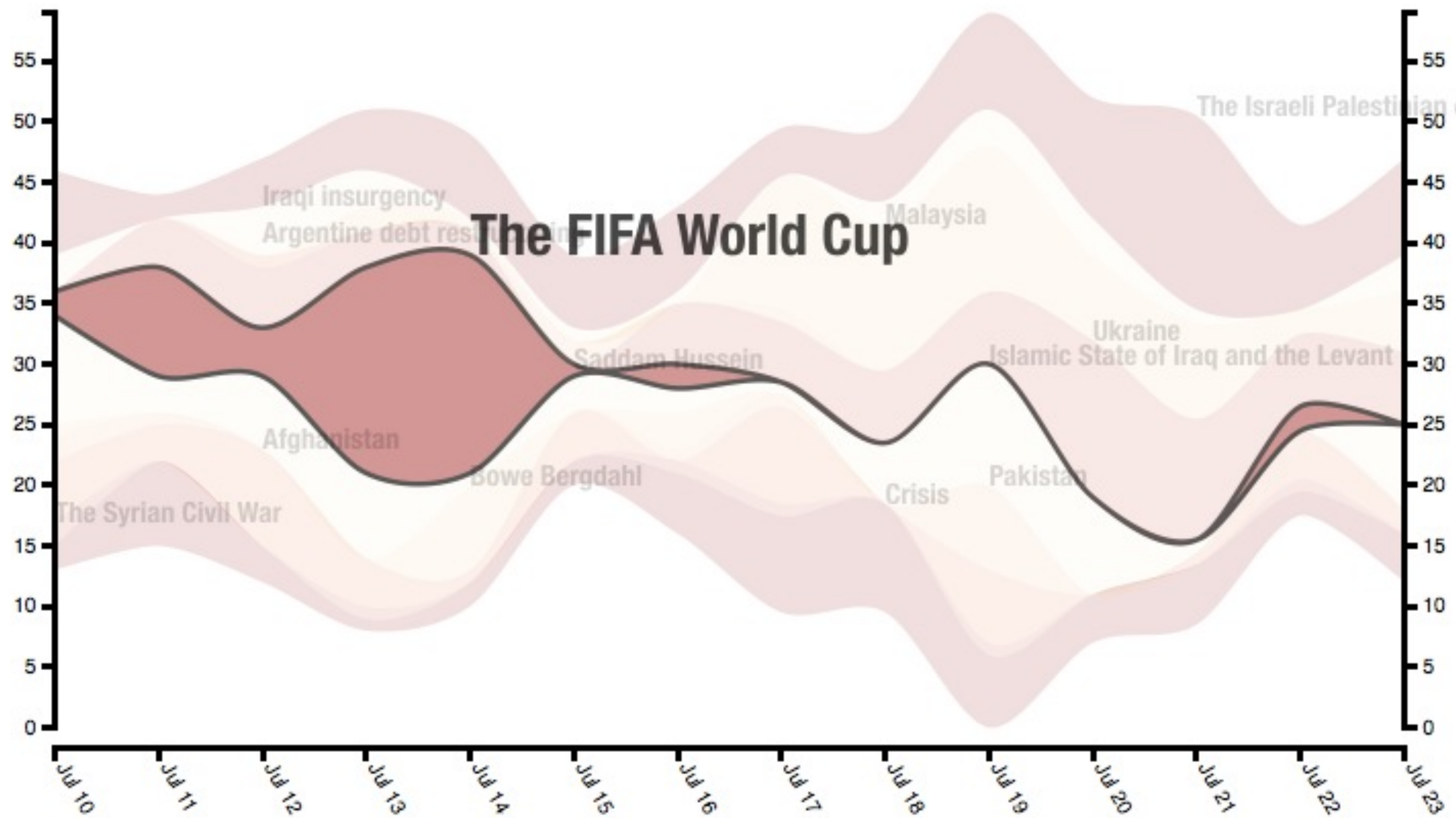
RATING	EXPLANATION
5	Title is one of major topics in Article. Article gives a particular attention to Title.
4	Part of Article deals with Title. Article makes a clear reference to Title.
3	Part of Title has some relevance to a dominant theme of Article. Example: Title 'European Tax System' is partially relevant to an article discussing US Tax System.
2	Article makes a reference to part of Title.
1	Title has no relevance to Article, in whatever way.

LANGUAGE	RATING	#instances
ENGLISH	4.63	97
JAPANESE	4.41	92

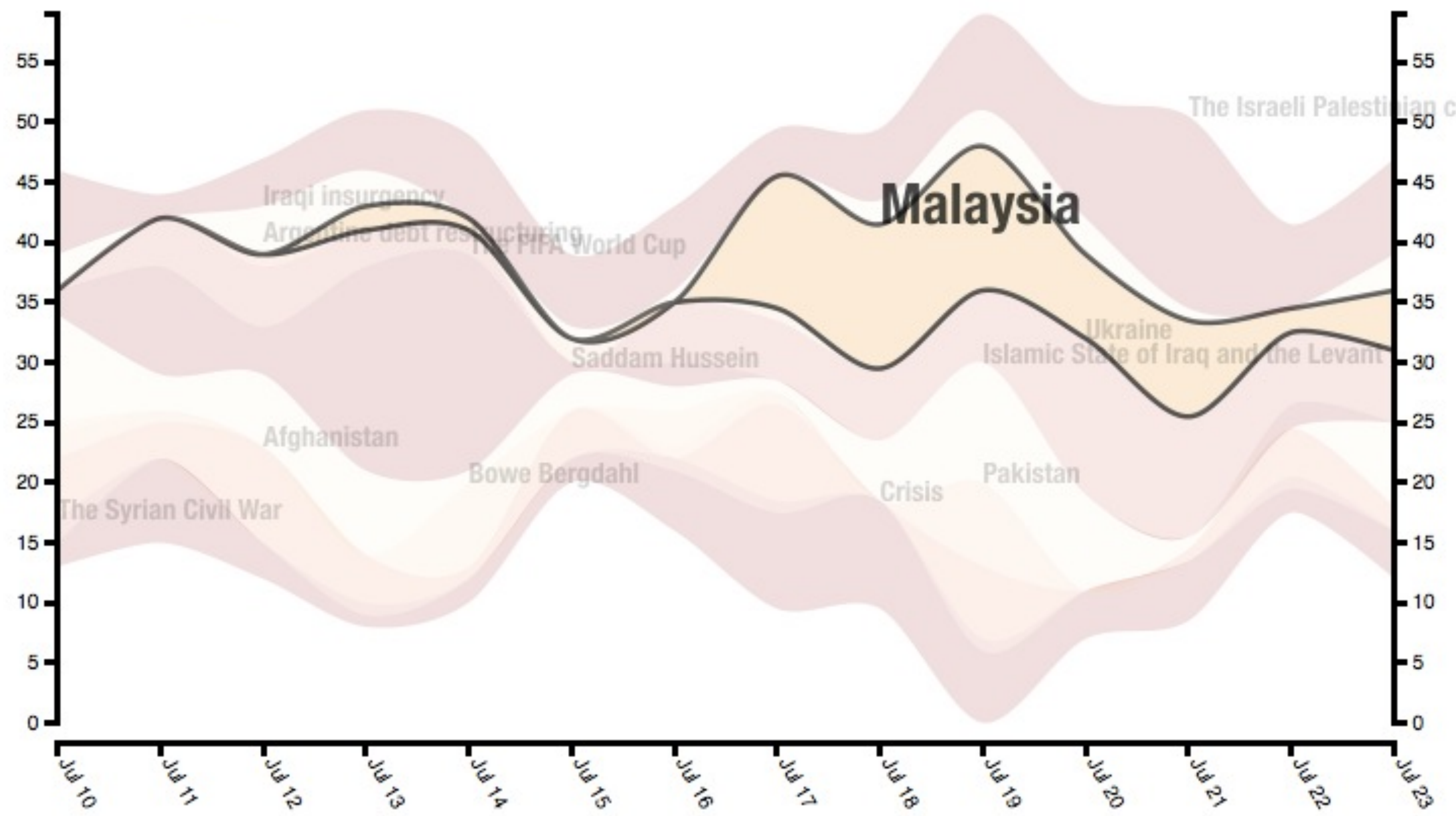
Trending Topics in the US Online News Media



Trending Topics in the US Online News Media



Trending Topics in the US Online News Media



Evaluation Metric: ROUGE-W

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

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

Text Rank vs. WikiLabel

	TRANK	RM_0	RM_1	RM_1/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

Solution to Problem

Frequencies of (manually assigned) topic descriptors that appeared in the New York Times from June to December, 2013.