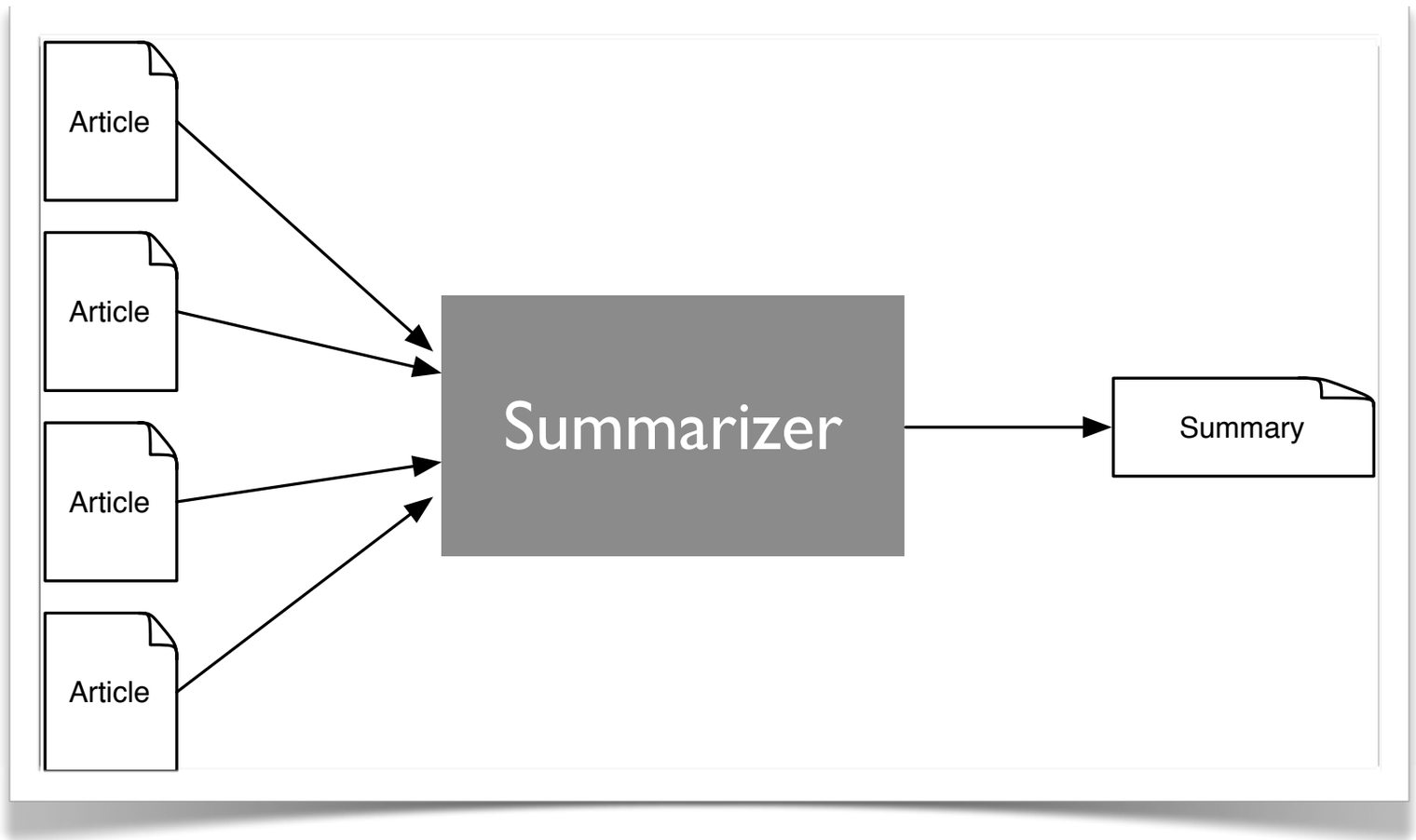
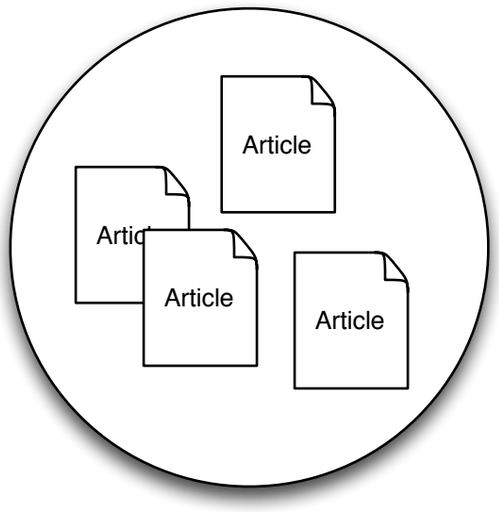
Exploiting Category-Specific Information for Multi-Document Summarization

Jun-Ping Ng Praveen Bysani Ziheng Lin
Min-Yen Kan Chew-Lim Tan

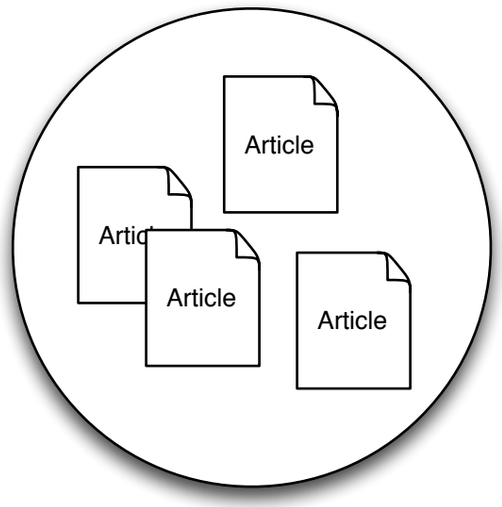
Multi-Document Summarization



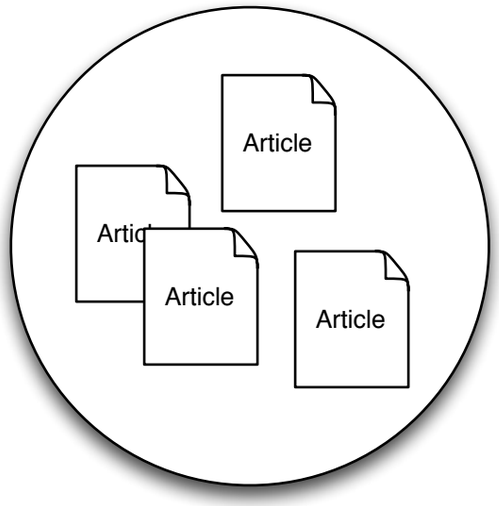
A Realistic Scenario



2012 US Presidential Election

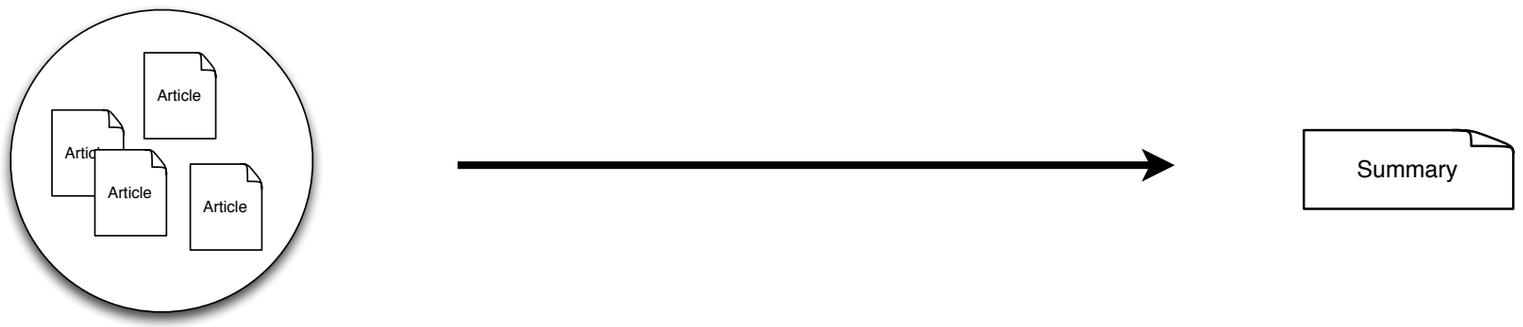


iPad Mini Release



Hurricane Sandy

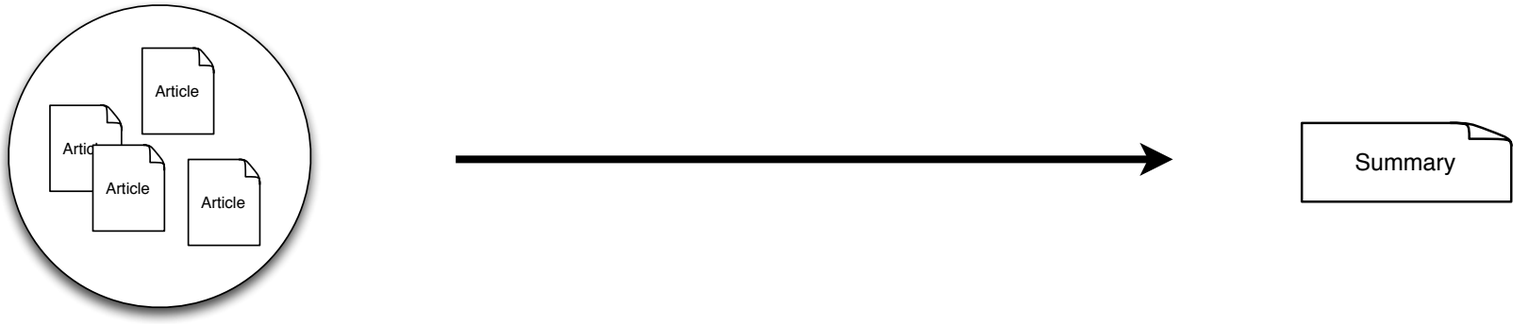
Typical Summarizers



2012 US Presidential Election

Typical Summarizers

2012 US Presidential Election

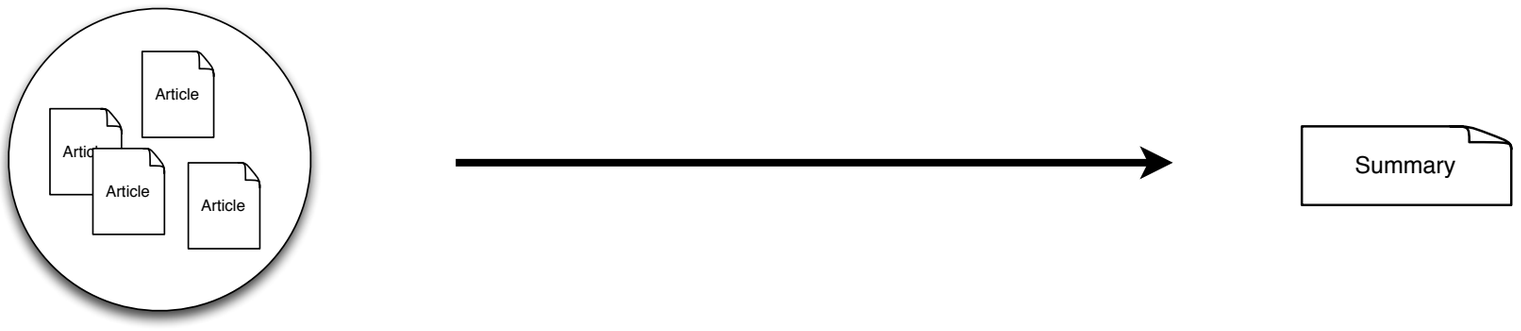


Hurricane Sandy

Typical Summarizers

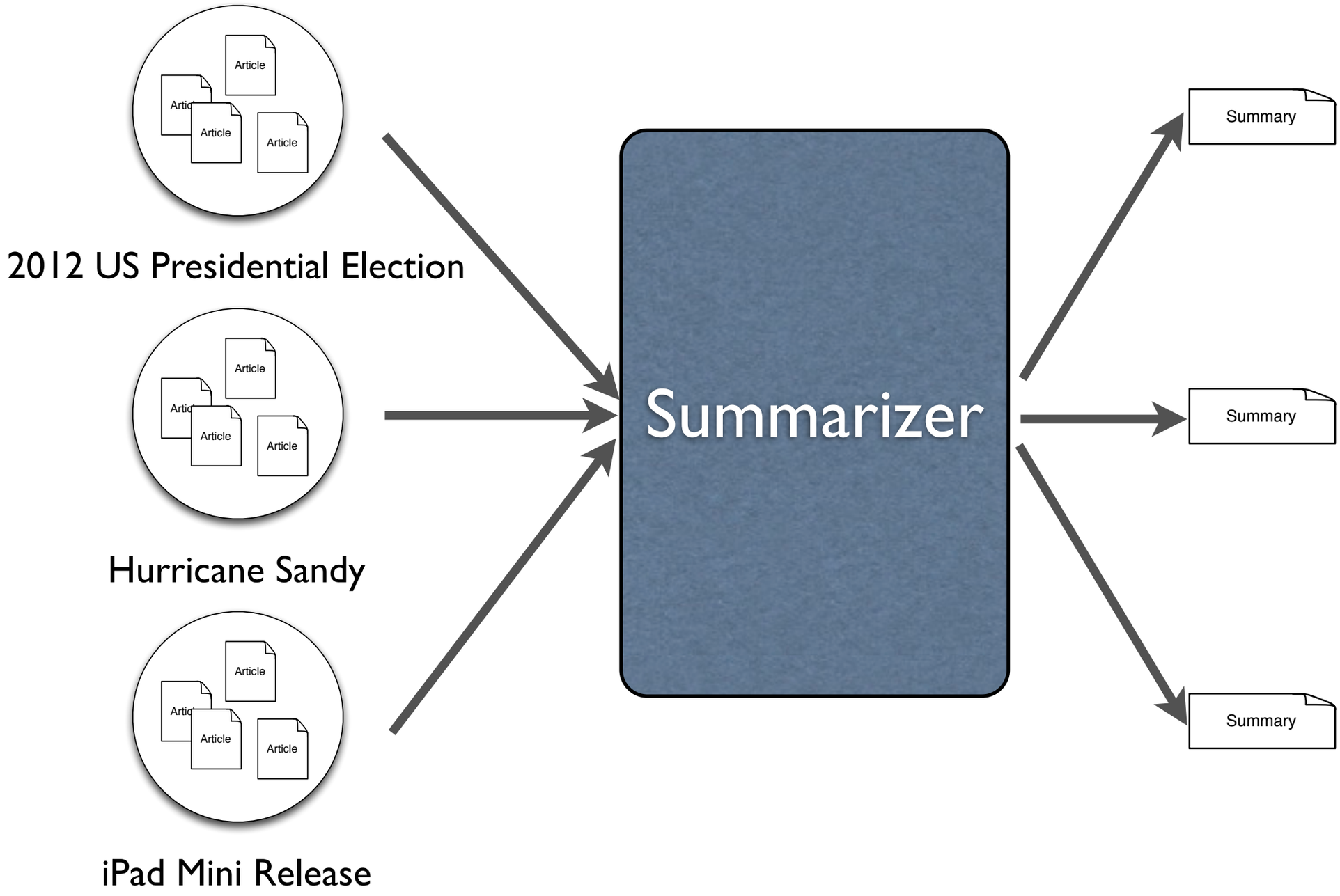
2012 US Presidential Election

Hurricane Sandy



iPad Mini Release

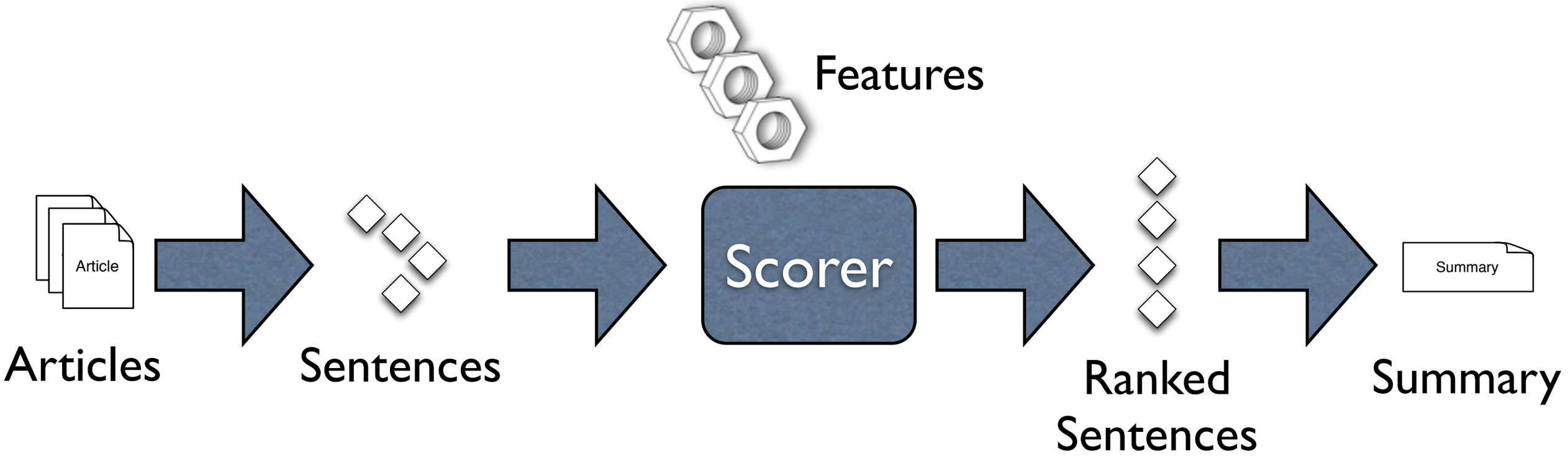
Key Insight


Article
Article
Article
2012 US Presidential Election
Article
Article
Article
Hurricane Sandy
Article
Article
Article
iPad Mini Release
Summarizer
Summary
Summary
Summary

Outline

- Baseline Summarization Pipeline
- Category-Specific Features
- Experiments

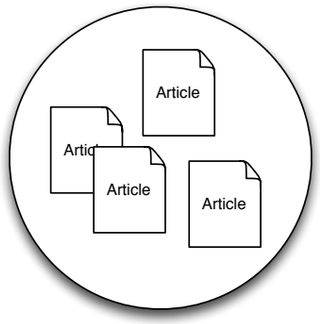
Pipeline


Features
Article
Scorer
Summary
Articles
Sentences
Ranked Sentences
Summary

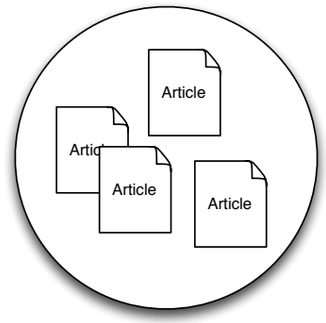
Generic Features

- Sentence position
- Sentence length
- Interpolated N-gram Document Frequency

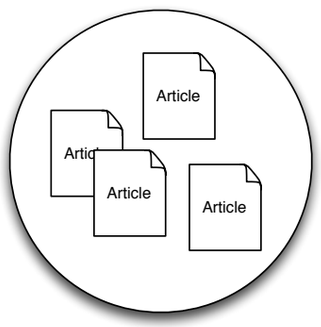
Category-Specific Features



2012 US Presidential Election



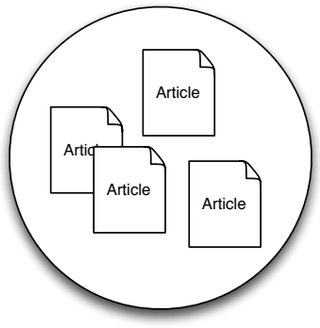
Hurricane Sandy



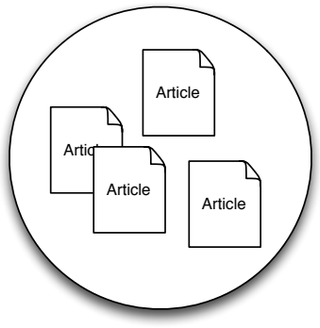
iPad Mini Release

9

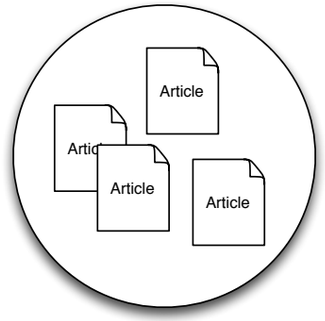
Category-Specific Features



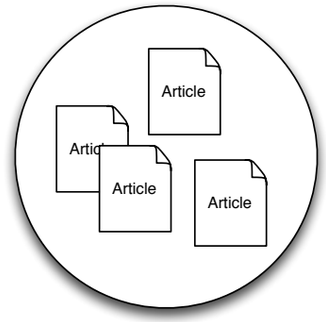
2012 US Presidential Election



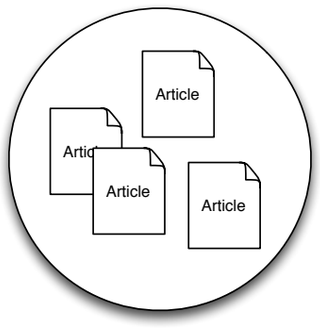
2011 UK Election



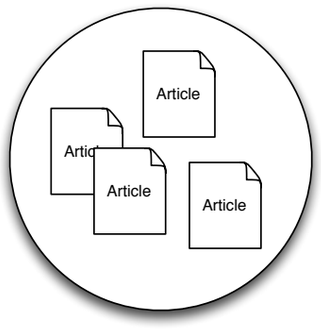
2012 Mynamar Earthquake



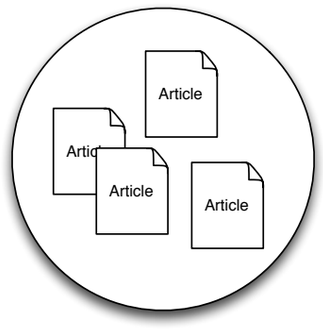
Hurricane Sandy



2012 Indian Assembly Elections

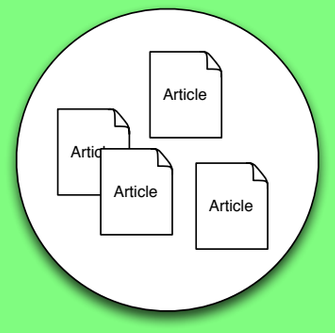


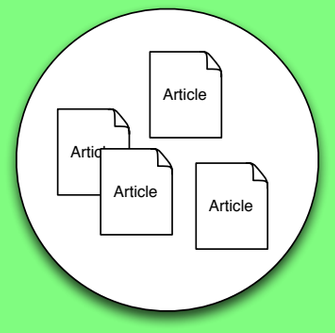
iPad Mini Release



Nokia Lumia 920

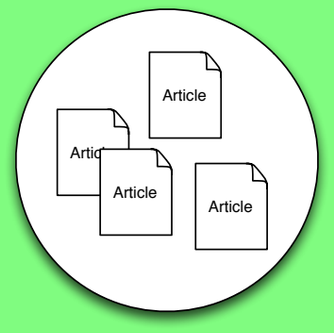
Category-Specific Features



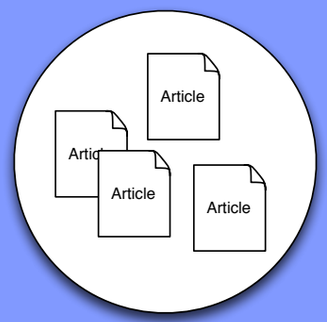


2012 US Presidential Election

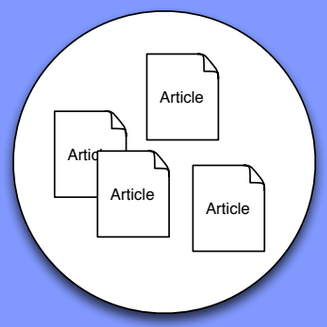
2011 UK Election



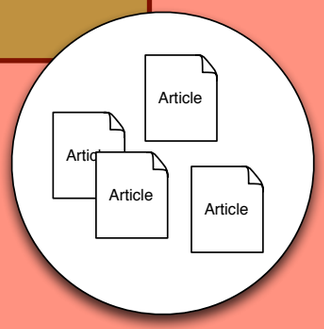
2012 Indian Assembly Elections



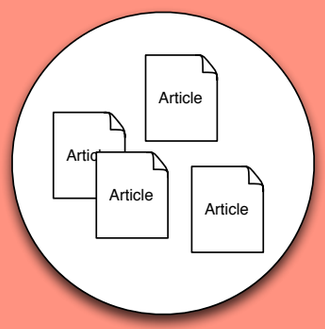
Hurricane Sandy



2012 Mynamar Earthquake



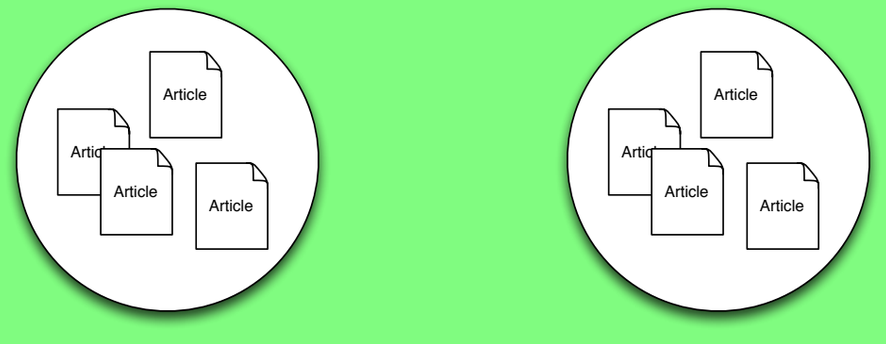
iPad Mini Release



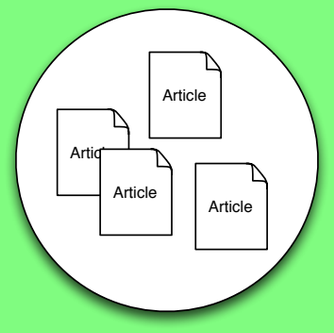
Nokia Lumia 920

Category-Specific Features

Disasters

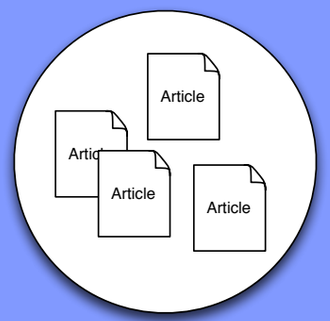


2012 US Presidential Election 2011 UK Election

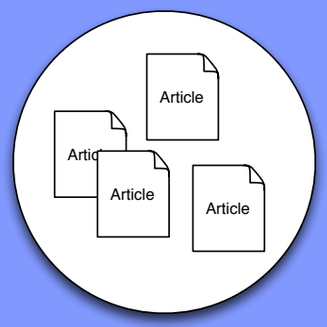


2012 Indian Assembly Elections

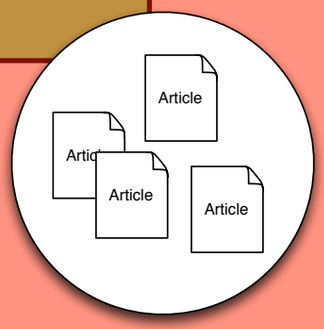
Political News



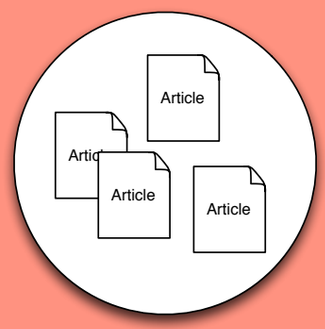
Hurricane Sandy



2012 Mynamar Earthquake



iPad Mini Release



Nokia Lumia 920

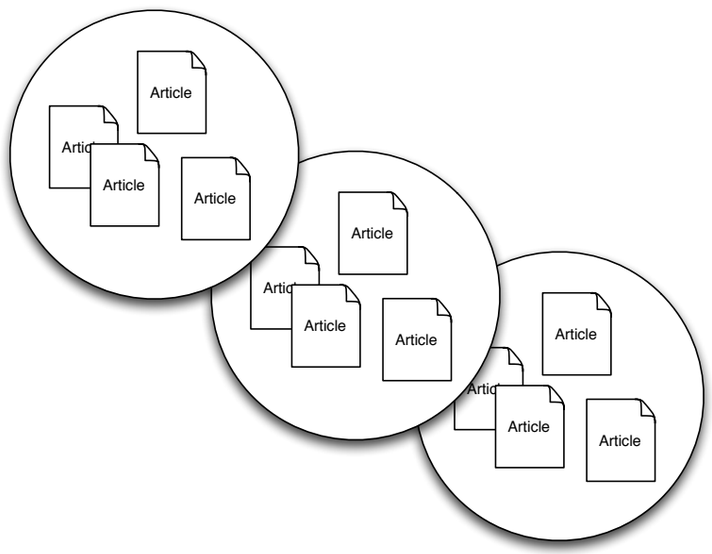
Technology

Data Set

- Made use of dataset from guided summarization task of TAC 2010/2011
- Sets of articles to be summarized grouped into topics
- Topics are grouped into categories

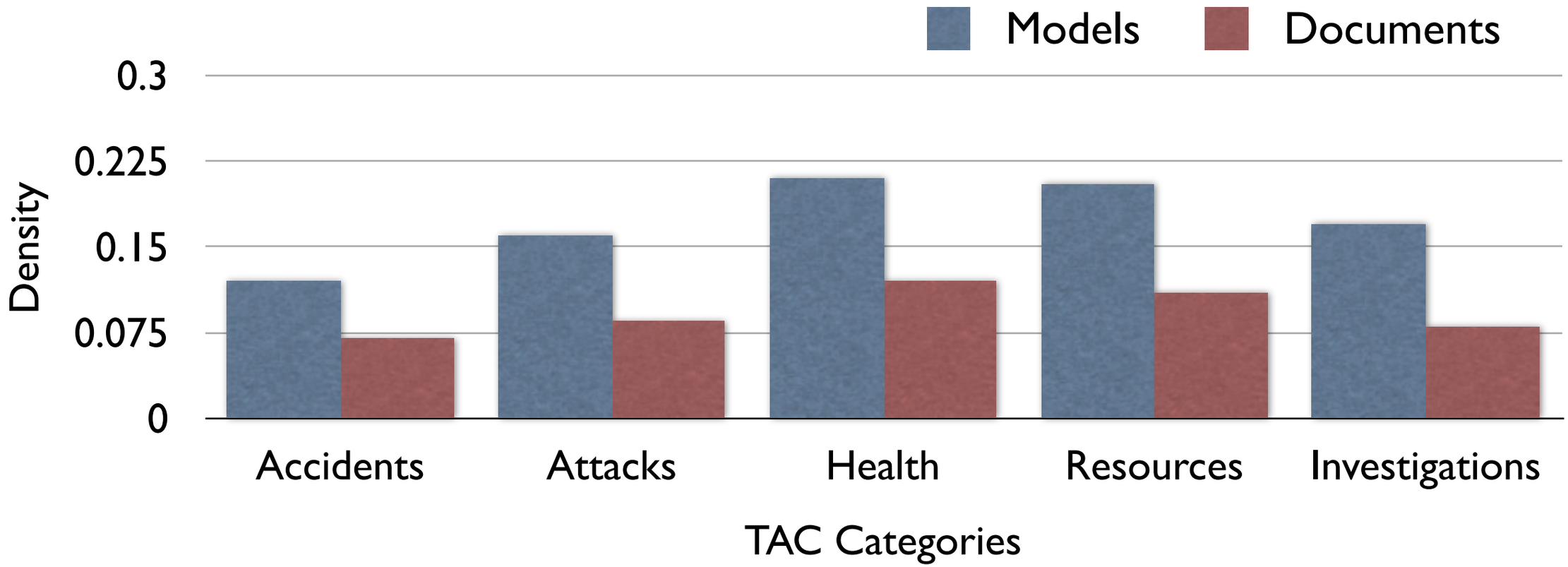
TAC Categories

- Accidents and Natural Disasters
- Attacks
- Health and Safety
- Endangered Resources
- Investigations and Trials



Justifying Our Intuition

- Words with high log-likelihood ratios are used more frequently inside model summaries



Category Relevance (CRS)

- **Importance** of a **word** with respect to a **category**

$$CRS_c(s) = \frac{\sum_{w \in s} (\beta \times TLF_c(w) + (1 - \beta) \times DLF_c(w))}{|s|}$$

Topic Frequency Score

Document Frequency Score

Category KLD (CKLD)

- Measure of **divergence** of probability **distribution** of a word in a **category**, with respect to the whole **collection**

$$CKLD_c(s) = \sum_{w \in s} \left(p_c(w) \times log \frac{p_c(w)}{p_C(w)} \right)$$

Probability of word in category

Probability of word in collection

Top Words

CRS	CKLD
official	crane
people	bridge
report	construction
news	java
accident	people

For articles on “Accidents”

Summarization Results

- Category-specific features give significant performance enhancement

System	ROUGE-2	ROUGE-4
Generic + CRS + CKLD	0.13796	0.16808
Generic + CRS	0.13702	0.16788
Generic + CKLD	0.13525	0.16649
CLASSY	0.12780	0.15812
POLYCOM	0.12269	0.15974

✣ ROUGE measures on TAC 2011 testing set

Less than Perfect Categorization

- In more realistic scenarios, perfect categorization is hard to come by
- Would our proposed CSI still be effective if we don't have perfect categorization?

Categorization Experiments

- Perfect clustering information is not essential

System	ROUGE-2	p-value
* Perfect Categories	0.13796	-
EM Clustering	0.13647	0.154
X-Means Clustering	0.13546	0.117
K-Means Clustering	0.13569	0.365

✣ Scores reported for 5 clusters

Analysis

Baseline

- The death toll could rise as thousands are still buried in debris and many are reported missing.

- Therefore, the relevant sectors and personnel should pay attention to disaster prevention.

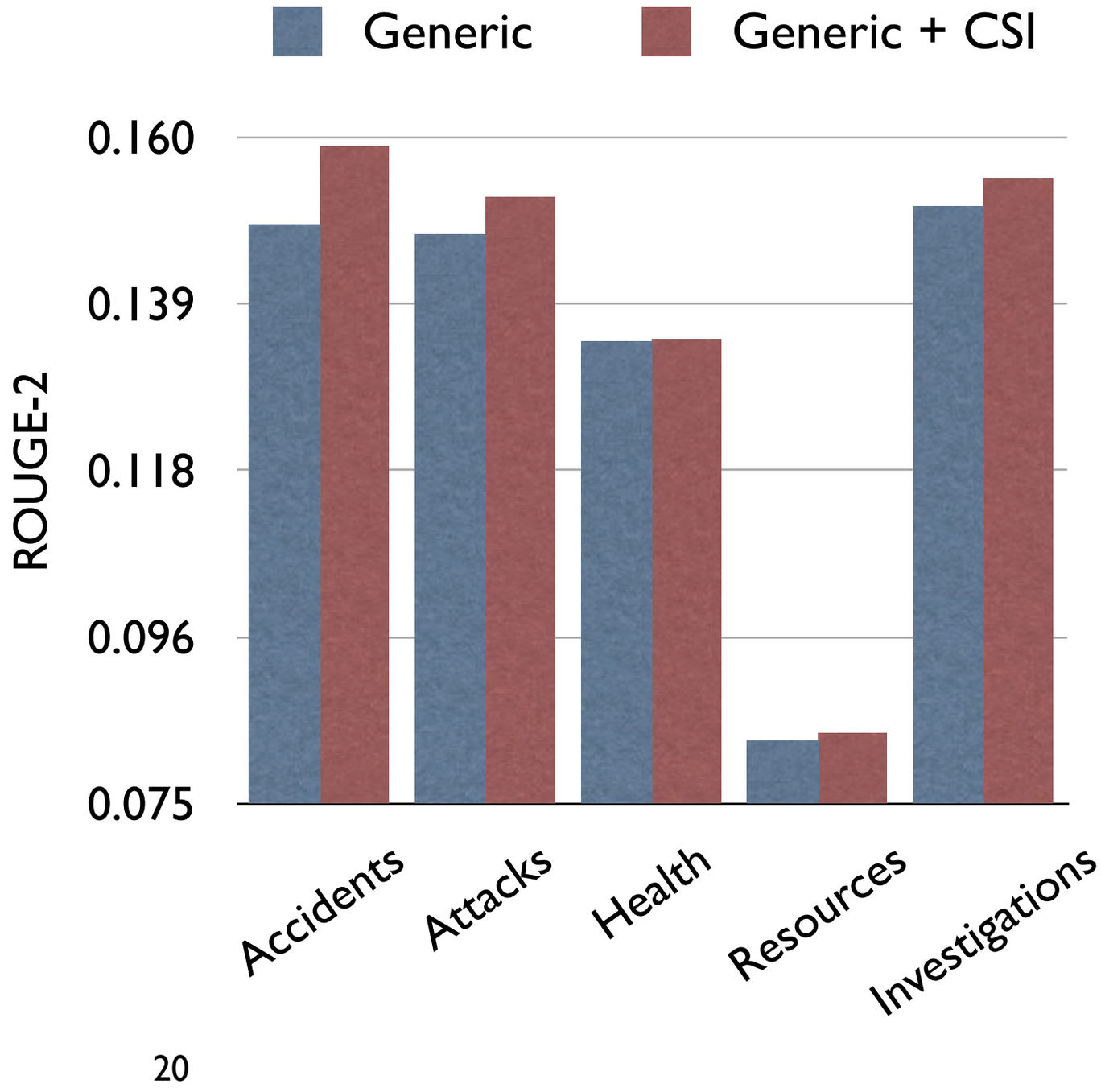
Generic + CRS + CKLD

+ Chinese authorities did not detect any warning signs ahead of Monday's earthquake that killed more than 8,600 people.

+ Xinhua said 8,533 people had died in Sichuan alone, citing the local government.

Varying Difficulties

- Some categories are easier to improve on
- Category-specific features help more with the WHO, WHAT, WHEN, WHERE



Round Up

- Category-specific information (CSI) can help multi-document summarization
- Effective in improving content selection
- Robust and is useful even when less than ideal categorization results are available

Thank you!

Questions?

You can download the system here:
http://wing.comp.nus.edu.sg/downloads/swing/