There is much more you can do with News than just reading them

Mouna Kacimi
KRDB Summer Seminar
31 July 2020
News are Important, we all know that...

They influence the way we think

They influence the way we behave
Why such diversity?

Different sources

Selective reading

Different versions

Increasing amount of information
How can we solve these issues

• Enrich information sources to have a wider view about news events

• Summarize news to have a quick access to information

• Analyze the effect of News on our behavior
In this talk

- News Enrichment
- News Information Extraction
- News Effect on Stock Markets

Summarization
The Article talks about

-Donald Trump warning Americans that coronavirus will get worse

-His previous policy of ignoring the coronavirus problem and pretending that US defeated it.

Published on July 21, 2020

Donald Trump, after spending months declaring the United States had defeated the coronavirus and vowing it soon will "disappear," on Tuesday warned Americans the outbreak that has killed at least 141,000 in the United States is likely to grow more dire.
Donald Trump, after spending months declaring the United States had defeated the coronavirus and vowing it soon will "disappear," on Tuesday warned Americans the outbreak that has killed at least 141,000 in the United States is likely to grow more dire.
Problem

Find news articles that match user interests

User

described by his comments

Latent Aspects
- China & WHO
- Russian disinformation

No matching

News Article

described by its content

News Aspects
- Donald Trump warning Americans

Problem

Find news articles that match user interests

User

described by his comments

Latent Aspects
- China & WHO
- Russian disinformation

No matching

News Article

described by its content

News Aspects
- Donald Trump warning Americans
Fill the gap between Comments & Articles

• Enrich news articles with comments

Content of a News Article

Content provided by each user

Both Articles and Users are described by News Aspects and Latent Aspects
Challenges

- How to enrich news articles with comments
  - Use all comments? No because they can be noisy
  - Select a subset of comments, but how?
- How aspects can be modeled and how to extract them from comments
Comments Selection

Comment the news article
Answer Comment

We use Page Rank Algorithm on the Tree Structure of Comments

\[ h(C) = (1 - m) + m \left( \frac{h(T_1)}{L(T_1)} + \ldots + \frac{h(T_n)}{L(T_n)} \right) \]
Comments Diversification

- Comments can be redundant (same idea)

- Two comments can describe the same aspects with 2 different writing style, so they are wrongly detected as diverse

- Use our opinion diversification algorithm proposed in [1]

Aspect Extraction

- Extract all words (tfidf)
- Construct phrases from the set of all words
- Phrases are constructed with an incrementing size 1,2,3,...
- Select meaningful phrases based on the correlation of their contained words (PMI)

Example of Generated Aspects

Trump, Clinton, Vote, Supporters, Business, Election, Establishment, President, crazies, Stupid Trump, Good Trump, Crazy Trump, Donald Trump vote, Donald Trump supporters, Donald Trump business, Donald Trump Establishment, Donald Trump president, Donald Trump crazies, Donald Trump trickster, Trump international policy
## Dataset Statistics

<table>
<thead>
<tr>
<th></th>
<th>Seeds from CNN</th>
<th>Seeds from The Telegraph</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#articles</td>
<td>#comments</td>
</tr>
<tr>
<td>CNN</td>
<td>41,245</td>
<td>12,056,789</td>
</tr>
<tr>
<td>Telegraph</td>
<td>1,908</td>
<td>1,257,645</td>
</tr>
<tr>
<td>Independent</td>
<td>1,412</td>
<td>987,437</td>
</tr>
<tr>
<td>Al-Jazeera</td>
<td>801</td>
<td>102,254</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Seeds from The INDEPENDENT</th>
<th>Seeds from Al-Jazeera</th>
</tr>
</thead>
<tbody>
<tr>
<td>#articles</td>
<td>#comments</td>
</tr>
<tr>
<td>CNN</td>
<td>528</td>
</tr>
<tr>
<td>Telegraph</td>
<td>23,272</td>
</tr>
<tr>
<td>Independent</td>
<td>27,012</td>
</tr>
<tr>
<td>Al-Jazeera</td>
<td>303</td>
</tr>
</tbody>
</table>
## Overall Precision of Our Approach

<table>
<thead>
<tr>
<th></th>
<th>P@1</th>
<th>P@3</th>
<th>P@5</th>
<th>P@10</th>
<th>P@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoEnrich</td>
<td>0.424</td>
<td>0.494</td>
<td>0.481</td>
<td>0.513</td>
<td>0.540</td>
</tr>
<tr>
<td>AllComments</td>
<td>0.393</td>
<td>0.474</td>
<td>0.445</td>
<td>0.453</td>
<td>0.503</td>
</tr>
<tr>
<td>Authority</td>
<td>0.454</td>
<td>0.535</td>
<td>0.530</td>
<td>0.550</td>
<td>0.565</td>
</tr>
<tr>
<td>Diversity</td>
<td>0.575</td>
<td>0.646</td>
<td>0.654</td>
<td>0.640</td>
<td>0.607</td>
</tr>
</tbody>
</table>

What News Articles Can Tell?

News events are reported by various sources.

An event can have several **facets:** reasons, consequences, time, location...

**How the firing was announced**

**Why Tillerson was fired**

**What is next?**
Why News Event Facets are important?

They enhance

- Information aggregation is facilitated
- Fluency can be improved
- Personalized

---

President Donald Trump announced that he had fired Secretary of State Rex Tillerson on Tuesday.

Trump fired Mr. Tillerson because he had defended the Iran Nuclear Deal.
Limitations of the State of the Art

- Open Information Extraction Tools (e.g., ClausIE, OLLIE)
  - Missing n-tuples,
  - coarse-grained argument
  - Trump announced that he had fired Mr. Tillerson on Tuesday because he had defended the Iran Nuclear Deal.

- Redundancies
  - No nested triples
  - Do not label any argument

- Semantic Role Labeling (SRL) (e.g., Illinoise, SEMAFOR)
  - Label some facets
US President Donald Trump announced that he had fired Secretary of State Rex Tillerson on Tuesday because he had defended the Iran Nuclear Deal.

1. Donald Trump; announced; _: facet
   • that; #2; (Other/Details)

2. he<ref#1s>; had fired; Rex Tillerson;
   facets
   • on; Tuesday; (Temporal)
   • because; #3; (Reason)

3. He<ref#2o>; defended; the Iran Nuclear Deal;

4. Donald Trump<ref#1s>; <is>; President;

5. Rex Tillerson<ref#1s>; <is>; Secretary of State;
Donald Trump; fired; Rex Tillerson; on; Tuesday; and appointed; Mike Pompeo; to replace; him <ref#1o>; 1.

2. Donald Trump<ref#1s>; <is>; US President
Learning Model for Facet Labeling:

- News Enrichment
- News Information Extraction
- News Effect on Stock Markets

Textual features via phrase embedding*

Type features via NER and SEMAFOR

Structural features:
- Word indexes
- Dep. graph & edges
- POS tagging
- ...

Multinomial logistic regression

~7000 instances
News Enrichment

News Information Extraction

News Effect on Stock Markets

- Compact graph representation
- Entity coreference
- Predicate Coreference

Summary:
US President Donald Trump announced that he had fired Secretary of State Rex Tillerson on Tuesday morning via Twitter and that he would replace him with CIA Director Mike Pompeo to have a new team in place. Mr. Tillerson learned he had been fired when a top aide showed him a tweet from Mr. Trump announcing the change. He was unaware of the reason of his firing. White House chief of staff John F. Kelly called Tillerson on Friday to warn him about the firing.
Summary Generation

- Rank facts
- Specify Depth D and number of facets F
- Proceed with Greedy Selection
Summary Generation

- Follow the order of facts in the original text
- Group facts with the same predecessor in the same sentence
- Use coreference for reduce redundancy.

**Summary:**
US President Donald Trump announced that he had fired Secretary of State Rex Tillerson on Tuesday morning via Twitter and that he would replace him with CIA Director Mike Pompeo to have a new team in place. Mr. Tillerson learned he had been fired when a top aide showed him a tweet from Mr. Trump announcing the change. He was unaware of the reason of his firing. White House Chief of Staff John F. Kelly called Tillerson on Friday to warn him about the firing.
### Experiments

#### News Enrichment

#### News Information Extraction

#### News Effect on Stock Markets

**TABLE IV**

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
<th>PYRAMID</th>
<th>LQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_1$</td>
<td>$F_1$</td>
<td>$F_1$</td>
<td>$T_{0.6}$</td>
<td>$T_{0.65}$</td>
</tr>
<tr>
<td>Peng li (RI)</td>
<td>29.31</td>
<td>5.77</td>
<td>10.01</td>
<td>65.6</td>
<td>58.9</td>
</tr>
<tr>
<td>Bing et al.,</td>
<td>n/a</td>
<td>11.70</td>
<td>14.80</td>
<td>90.5</td>
<td>79.3</td>
</tr>
<tr>
<td>GRANULA</td>
<td><strong>43.17</strong></td>
<td><strong>12.14</strong></td>
<td><strong>16.52</strong></td>
<td><strong>90.5</strong></td>
<td><strong>79.6</strong></td>
</tr>
</tbody>
</table>

**Comparison with the baselines over TAC’11**

**TABLE V**

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
<th>PYRAMID</th>
<th>LQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_1$</td>
<td>$F_1$</td>
<td>$F_1$</td>
<td>$T_{0.6}$</td>
<td>$T_{0.65}$</td>
</tr>
<tr>
<td>17</td>
<td>43.21</td>
<td>12.99</td>
<td>15.97</td>
<td>86.0</td>
<td>74.1</td>
</tr>
<tr>
<td>43</td>
<td><strong>43.82</strong></td>
<td><strong>13.45</strong></td>
<td><strong>16.52</strong></td>
<td><strong>87.5</strong></td>
<td><strong>75.8</strong></td>
</tr>
<tr>
<td>22</td>
<td>42.79</td>
<td>11.45</td>
<td>15.92</td>
<td>87.8</td>
<td>77.5</td>
</tr>
<tr>
<td>GRANULA</td>
<td>43.17</td>
<td>12.14</td>
<td><strong>16.52</strong></td>
<td><strong>90.5</strong></td>
<td><strong>79.6</strong></td>
</tr>
</tbody>
</table>

**Comparison with the top-3 extractive methods of TAC’11**
How News can affect our trading behavior in stock markets?
What type of data we deal with?

Prices

Trading Volume
How to combine both types of data to predict stock prices?
Challenges

News can be many, not all useful for predicting stock prices

What types of News we should look at? Specific to the stock market of interest or general news?

What part of the text we should focus on? Title, keyworks, topic, actions, entities, all the text?

Which news have short or long-lasting effects?
Proposal: redefine what a news event

Event: any piece of news that makes price change on a specific stock market

Single prices

Patterns
Proposal: exploit our multi-granular summarizer

Summarize a set of news articles about the same event in one single text.

Produce summaries with different granularities ranking from the general topic detailed facts and use content for feature extraction.

StuffIE
Proposal: use LSTM networks for stock price prediction

- Long short-term memory RNN Networks
- Typically used for times series data
- Avoid long term dependency problem
Thank you...
Proposal: exploit our multi-granular summarizer

Summarize a set of news articles about the same event in one single text

Produce summaries with different granularities ranking from the general topic detailed facts and use content for feature extraction
Facet Labeling

He had fired Secretary of State Rex Tillerson; on Tuesday.

Facet labels: where to obtain them?
Generating training data for facet labeling
Learning model for facet labeling
Facet labels: where to obtain good labels?

Facets are connected by prepositions and conjunctions

<table>
<thead>
<tr>
<th>Prep &amp; conj</th>
<th>No</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>for</td>
<td>1.</td>
<td>Agent</td>
</tr>
<tr>
<td>on</td>
<td>2.</td>
<td>Beneficiary</td>
</tr>
<tr>
<td>because</td>
<td>3.</td>
<td>Cause</td>
</tr>
<tr>
<td>while</td>
<td>4.</td>
<td>Location</td>
</tr>
<tr>
<td>by</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>34.</td>
<td>Temporal</td>
</tr>
<tr>
<td></td>
<td>35.</td>
<td>Topic</td>
</tr>
</tbody>
</table>

Illinois SRL coverage check

Wiktionary The free dictionary

+additions
Distant Supervision: Building Training Data

Sentence: “The fleet is on the American coast.”
Facet conn: on
Gloss: “At or near; adjacent to”

Sentence: “He left the house on Monday.”
Facet conn: on
Gloss: “Some time during the day of”

Sentence: “The fleet is on the American coast.”
Facet conn: on
Label: Location

Sentence: “He left the house on Monday.”
Facet conn: on
Label: Temporal
Enrich The Labels

Facet label: “Temporal”

Synonymous / related words


No | Label
---|---
1. | Agent
2. | Beneficiary
3. | Cause
4. | Location
34. | Temporal
35. | Topic

No | Context-enriched label
---|---
1. | <Agent, Instigator, Actor, Initiator,…>
2. | <Beneficiary, Recipient, Receiver, …>
3. | <Cause, Reason, Grounds, …>
4. | <Location, Region, Space, …>
34. | <Temporal, Time, Period, Day, …>
35. | <Topic, Subject, Theme, …>
Map Glosses to Similar Labels!

“Some time during the day of”

Word tokenization

\[\text{<Some, time, during, the, day, of>}\]

Count overlap

No | Context-enriched label
---|-----------------------
1  | \(<\text{Agent, Instigator, Actor, Initiator,}…\)>  
2  | \(<\text{Beneficiary, Recipient, Receiver,}…\)>  
3  | \(<\text{Cause, Reason, Grounds,}…\)>  
4  | \(<\text{Location, Region, Space,}…\)>  
...|
34 | \(<\text{Temporal, Time, Period, Day,}…\)>  
35 | \(<\text{Topic, Subject, Theme,}…\)>  

<table>
<thead>
<tr>
<th>Overlaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Overlap tiebreakers
- Dependency edge distance
- Word distance
- Word2vec + WordNet::Similarity averages

We evaluated the mapping on a ground truth of 100 pairs and reached 93% of precision.
Learning Model for Facet Labeling: Summing up

Crawling ~7000 instances

Textual features via phrase embedding*

Type features via NER and SEMAFOR

Structural features:
- Word indexes
- Dep. graph & edges
- POS tagging
- ...

Multinomial logistic regression
Experiments: Questions about StuffIE

• How good is StuffIE’s facet labeling model to predict its own training data?
  • Which features are the best?

• How does StuffIE’s facet labeling perform compared to Illinois SRL and SEMAFOR SRL?
Experiments: Questions about StuffIE

• How many triples and facets does StuffIE extract, compared to ClausIE and OLLIE?
Experiments and Results: Triple & Facet Extraction

**Dataset:** DUC’04 news dataset with ~20,000 sentences

<table>
<thead>
<tr>
<th>Method</th>
<th>Triples</th>
<th>Facets</th>
<th>Nested</th>
<th>Miss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClausIE</td>
<td>76906</td>
<td>105840</td>
<td>16381</td>
<td>158</td>
</tr>
<tr>
<td>OLLIE</td>
<td>70862</td>
<td>69781</td>
<td>161</td>
<td>642</td>
</tr>
<tr>
<td>StuffIE</td>
<td>83265</td>
<td>107210</td>
<td>22740</td>
<td>0</td>
</tr>
</tbody>
</table>

*Miss = missing extraction

**Metrics:** The number of relations, “facets”, “nested relations”, and missing extractions.

**Note:** “Facets” for ClauSIE and OLLIE are counted by the number of arguments.
“Nested relations” are counted by redundant extractions.