DISTRIBUTED NON-PARAMETRIC REPRESENTATIONS FOR VITAL FILTERING:
UW @ TREC KBA 2014

IGNACIO CANO - SAMEER SINGH - CARLOS GUESTRIN

TEXT RETRIEVAL CONFERENCE - NOVEMBER 2014
VITAL FILTERING (VF)

entities

documents

each doc
We use two classifiers in cascade.

VITAL FILTERING (VF)

R-NR

if R

V-NV

if NR/UNK

NON-REFERENT

VITAL

NON-VITAL
Agenda

- METHOD
- EXPERIMENTS
- ACCELERATE & CREATE
- CONCLUSION
METHOD

• We need to predict whether a document contains novel information for an entity

• If we can “summarize” the contents seen for an entity

• We can compare the document to the entity’s summary

  • If similar, we have probably seen the information

  • If dissimilar, it is novel -> predict VITAL

• How can we accurately and efficiently “summarize” all the documents that we have seen for an entity?
**BAG OF WORDS**

+ word counts are simple to implement, often work quite well

- sparsity

“Barack was elected president today”

“Obama has won the election”

- size of representation grows over time
WORD EMBEDDINGS

- Vector representations of words
- fixed length array of real values

\[ V_{france} = (0.002, 0.2, 0.4\ldots) \]

\[ V_{paris} = (0.005, 0.18, 0.43\ldots) \]
So, it comes as good news that state officials are making significant progress in determining which chemicals pollute Puget Sound and in identifying where they come from. A report last week from the Department of Ecology casts a wide net over culprits. Most toxic chemicals are used in some way by all of us, said Ecology Director Ted Sturdevant. They are in our homes and gardens...
So, it comes as good news that state officials are making significant progress in determining which chemicals pollute Puget Sound and in identifying where they come from. A report last week from the Department of Ecology casts a wide net over culprits. Most toxic chemicals are used in some way by all of us, said Ecology Director Ted Sturdevant. They are in our homes and gardens...
Most toxic chemicals are used in some way by all of us, said Ecology Director Ted Sturdevant.
e = Ted Sturdevant

chemicals are used way

said Ecology Director Ted Sturdevant
DOCUMENT EMBEDDING

e = Ted Sturdevant

\[ W = [ \text{chemicals, are, used, way, said, ecology, director, ted_sturdevant} ] \]

\[ v_{\text{chemicals}} = f(\text{chemicals}) \]
\[ v_{\text{are}} = f(\text{are}) \quad v_{\text{said}} = f(\text{said}) \]
\[ v_{\text{used}} = f(\text{used}) \quad v_{\text{ecology}} = f(\text{ecology}) \]
\[ v_{\text{way}} = f(\text{way}) \quad v_{\text{director}} = f(\text{director}) \]

\[ v_d = \frac{1}{|W|} (v_{\text{chemicals}} + v_{\text{are}} + v_{\text{used}} + v_{\text{way}} + v_{\text{said}} + v_{\text{ecology}} + v_{\text{director}}) \]
METHOD

• So far we have a representation for documents.
• To summarize what we have seen of an entity
  \[ v_e = \frac{1}{|D_e|} \sum_{d \in D_e} v_d \]
• Compare embedding of doc to entity’s
• If close -> predict NON-VITAL
e=Obama

senator
e=Obama
e=Obama

president

president + senator

senator
MULTIPLE EMBEDDINGS FOR ENTITIES

• Use multiple embeddings to capture the different clusters of an entity.

• We propose an algorithm that:
  • assigns each document to a single cluster,
  • adaptively grows the number of clusters
  • is efficient
e = Obama
e=Obama

\begin{align*}
\text{senator} \\
\alpha
\end{align*}
e=Obama
$e=\text{Obama}$
e=Obama
METHOD

- Take embedding of document
- Compare to all cluster vectors of an entity
  - If close to ANY -> we have seen similar document
  - If far from ALL -> document contains novel info
METHOD

• Unfortunately this is not accurate

“Obama wins the presidential election” (2008)

“Obama wins his second presidential election” (2012)
METHOD

- Propose a staleness measure
  - low staleness -> novel, freshness, vital
  - high staleness -> old, stale, non-vital
e = Obama


Nov 4, 2008 - Barack Hussein Obama was elected the 44th president of the United States on Tuesday, as the country chose him as its first black chief executive. Monitor live results all night on The Times's Election 2008 Dashboard.

STALENESS

0

1

p(vital)

high
e=Obama

STALENESS

0

high

p(vital)

1

high
time

  - Nov 4, 2008 - Barack Hussein Obama was elected the 44th president of the United States on Tuesday, as the country chose him as its first black chief executive. Monitor live results all night on The Times's Election 2008 Dashboard.

  - Nov 4, 2008 - Barack Hussein Obama was elected the 44th president of the United States on Tuesday, as the country chose him as its first black chief executive. Monitor live results all night on The Times's Election 2008 Dashboard.

- **CNN Announces Obama as the Next President Elect** ...
  - www.youtube.com/watch?v=9dKAKf1bUE
  - Nov 5, 2008 - Uploaded by GettingtheTruth2
  - Lies upon lies coming from the shit hole which is obamas mouth. ... Election 2008 - United States Presidential Inauguration by Marc Schulz ...

- **American President: Barack Obama: Campaigns and ...**
  - millercenter.org/president/obama/essays/
  - Miller Center of Public Affairs - Obama's election to the Senate instantly made him the highest-ranking ..... His margin of victory was down slightly from 2008, making him the first President since ...
e=Obama

www.nytimes.com/2008/11/05campaign.html?.. - The New York Times - Nov 4, 2008 - Barack Hussein Obama was elected the 44th president of the United States on Tuesday, as the country chose him as its first black chief executive. ... Monitor live results all night on The Times's Election 2008 Dashboard.

www.nytimes.com/2008/11/05campaign.html?.. - The New York Times - Nov 4, 2008 - Barack Hussein Obama was elected the 44th president of the United States on Tuesday, as the country chose him as its first black chief executive. ... Monitor live results all night on The Times's Election 2008 Dashboard.

CNN Announces Obama as the Next President Elect ...
www.youtube.com/watch?v=9kKAKftbUE
Nov 5, 2008 - Uploaded by GettingTheTruth2
... Lies upon lies coming from the shit hole which is obamas mouth. ... Election 2008 - United States Presidential Inauguration by Marc Schulz ...

American President: Barack Obama: Campaigns and ...
millercenter.org/president/obama/essays/...
Obama's election to the Senate instantly made him the highest-ranking ... His margin of victory was down slightly from 2008, making him the first President since ...

Obama elected nation's 44th president - USATODAY.com
usatoday30.usatoday.com/news/election2008/president.htm...
USA Today - Nov 1, 2008 - Democrat Barack Obama has been elected the 44th president of the United ... making history by becoming the first African American to win the nation's highest office. ... Updated 11/5/2008 2:03 PM | Comment | Recommend, E-mail | Print ...

Presidential Election of 2008 - 270toWin.com
www.270towin.com/2008_Election/...
270towin.com - Results of the presidential election of 2008, won by Barack H. Obama with 365 electoral ... It was also the first time two sitting senators ran against each other.
e=Obama

www.nytimes.com/2008/11/05campaign.html?
Nov 4, 2008 - Barack Hussein Obama was elected the 44th president of the United States on Tuesday, as the country chose him as its first black chief executive. .... Monitor live results all night on The Times's Election 2008 Dashboard.

www.nytimes.com/2008/11/05campaign.html?
Nov 4, 2008 - Barack Hussein Obama was elected the 44th president of the United States on Tuesday, as the country chose him as its first black chief executive. .... Monitor live results all night on The Times's Election 2008 Dashboard.

CNN Announces Obama as the Next President Elect ...
www.youtube.com/watch?v=9dKAK11bUE
Nov 5, 2008 - Uploaded by GettingtheTruth2
Lies upon lies coming from the shit hole which is obamas mouth. ... Election 2008 - United States Presidential Inauguration by Marc Schulz ...

American President: Barack Obama: Campaigns and ...
millercenter.org/president/obama/essays/...
Obama's election to the Senate instantly made him the highest-ranking .... His margin of victory was down slightly from 2008, making him the first President since ...

Obama elected nation's 44th president - USATODAY.com
usatoday30.usatoday.com/news/.../election2008/president.ht...
USA Today ~ Nov 1, 2008 - Democrat Barack Obama has been elected the 44th president of the United ... making history by becoming the first African American to win the nation's highest office. ... Updated 11/5/2008 2:03 PM | Comment | Recommend, E-mail | Print ...

Presidential Election of 2008 - 270toWin.com
www.270towin.com/2008_Election/...
Results of the presidential election of 2008, won by Barack H. Obama with 365 electoral .... It was also the first time two sitting senators ran against each other.
e=Obama

www.nytimes.com/2008/11/05/elections/results.html
Nov 4, 2008 - Barack Hussein Obama was elected the 44th president of the United States on Tuesday, as the country chose him as its first black chief executive. ... Monitor live results all night on The Times's Election 2008 Dashboard.

CNN Announces Obama as the Next President Elect ...
www.youtube.com/watch?v=9DKA11tUE
Nov 5, 2008 - Uploaded by GettingTheTruth2
... Election 2008 - United States Presidential Inauguration by Marc Schulz ...

American President: Barack Obama: Campaigns and ...
millercenter.org/president/obama/essays/... - Miller Center of Public Affairs
... Obama's election to the Senate instantly made him the highest-ranking ... His margin of victory was down slightly from 2008, making him the first President since ...

Obama elected nation's 44th president - USATODAY.com
usatoday30.usatoday.com/news/election2008/president.ht... - USA Today
Nov 1, 2008 - Democrat Barack Obama has been elected the 44th president of the United ... making history by becoming the first African American to win the nation's highest office. ... Updated 11/8/2008 2:03 PM | Comment | Recommend, E-mail | Print ...

Presidential Election of 2008 - 270toWin.com
www.270towin.com/2008_Election/... - 270towin.com
Results of the presidential election of 2008, won by Barack H. Obama with 365 electoral ... It was also the first time two sitting senators ran against each other.

Election 2012: Obama re-elected president - NY Daily News
Nov 6, 2012 - President Barack Obama won re-election Tuesday night despite a ... that encumbered his first term and cramped the middle class dreams of ...
STALENESS

• Take embedding of document
• Compare to all cluster vectors of an entity
  • If far from ALL -> document contains novel info
  • If close to ANY and LOW staleness -> VITAL
  • If close to ANY and HIGH staleness -> NON-VITAL
Agenda

- VITAL FILTERING
- METHOD
- EXPERIMENTS
- ACCELERATE & CREATE
- CONCLUSION
FEATURES

• Baseline
  • Document Level
    • log length and source of the document
  • Document-Entity Level
    • number and position of occurrences of entities in the document, etc.
OUR FEATURES

- Embedding
  - embedding of document

- Clustering
  - min and average distances between embeddings of document and clusters

- Temporal
  - staleness of the entity
  - staleness of the assigned cluster
## Model Ablation Results

<table>
<thead>
<tr>
<th>MODEL</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.475</td>
<td>0.238</td>
<td>0.317</td>
</tr>
<tr>
<td>Baseline Multi-task</td>
<td>0.367</td>
<td>0.405</td>
<td>0.385</td>
</tr>
</tbody>
</table>
# Model Ablation Results

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.475</td>
<td>0.238</td>
<td>0.317</td>
</tr>
<tr>
<td>Baseline Multi-task</td>
<td>0.367</td>
<td>0.405</td>
<td>0.385</td>
</tr>
<tr>
<td>Embedding</td>
<td>0.449</td>
<td>0.376</td>
<td>0.409</td>
</tr>
</tbody>
</table>
## Model Ablation Results

<table>
<thead>
<tr>
<th>MODEL</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.475</td>
<td>0.238</td>
<td>0.317</td>
</tr>
<tr>
<td>Baseline Multi-task</td>
<td>0.367</td>
<td>0.405</td>
<td>0.385</td>
</tr>
<tr>
<td>Embedding</td>
<td>0.449</td>
<td>0.376</td>
<td>0.409</td>
</tr>
<tr>
<td>Embedding-Temporal</td>
<td>0.444</td>
<td>0.358</td>
<td>0.397</td>
</tr>
</tbody>
</table>
## Model Ablation Results

<table>
<thead>
<tr>
<th>MODEL</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.475</td>
<td>0.238</td>
<td>0.317</td>
</tr>
<tr>
<td>Baseline Multi-task</td>
<td>0.367</td>
<td>0.405</td>
<td>0.385</td>
</tr>
<tr>
<td>Embedding</td>
<td>0.449</td>
<td>0.376</td>
<td>0.409</td>
</tr>
<tr>
<td>Embedding-Temporal</td>
<td>0.444</td>
<td>0.358</td>
<td>0.397</td>
</tr>
<tr>
<td>Embedding-Clustering</td>
<td>0.439</td>
<td>0.372</td>
<td>0.403</td>
</tr>
</tbody>
</table>
### Model Ablation Results

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.475</td>
<td>0.238</td>
<td>0.317</td>
</tr>
<tr>
<td>Baseline Multi-task</td>
<td>0.367</td>
<td>0.405</td>
<td>0.385</td>
</tr>
<tr>
<td>Embedding</td>
<td>0.449</td>
<td>0.376</td>
<td>0.409</td>
</tr>
<tr>
<td>Embedding-Temporal</td>
<td>0.444</td>
<td>0.358</td>
<td>0.397</td>
</tr>
<tr>
<td>Embedding-Clustering</td>
<td>0.439</td>
<td>0.372</td>
<td>0.403</td>
</tr>
<tr>
<td>Embedding-Temporal-Clustering</td>
<td>0.452</td>
<td>0.378</td>
<td><strong>0.412</strong></td>
</tr>
</tbody>
</table>
OFFICIAL VITAL RESULTS

P = 2nd
F1 = 8th
UNOFFICIAL VITAL RESULTS

F1 = 2nd
Agenda

- VITAL FILTERING
- METHOD
- EXPERIMENTS
- ACCELERATE & CREATE
- CONCLUSION & FUTURE WORK
VISUALIZATION

• Browser-based visualization prototype with interactive widgets

• Mainly for debugging purposes
CLUSTER VIEW

legend / filter

active section

click opens wordcloud

time range filter
Agenda

• VITAL FILTERING
• METHOD
• EXPERIMENTS
• ACCELERATE & CREATE
• CONCLUSION
CONCLUSION

- Semi-supervised model for document filtering tasks
- Low dimensional vector based semantics is useful for VF
- Non-parametric clustering can capture multiple aspects of entities
- Simple decay based temporal dynamics indicate novelty
- Official results show we are 2nd in Precision and 3rd in SU
- Follow up analysis suggests we are 2nd in F1 as well