Global and Local Models for Multi-document Summarization

Pradipto Das and Rohini Srihari
SUNY Buffalo
TAC 2011, Gaithersburg, MD
Global and Local Models

Accidents and Natural Disasters

Attacks (Criminal/Terrorist)

Health and Safety

Endangered Resources

Investigations and Trials (Criminal/Legal/Other)

Local Models

Global Model

Multi-document summaries
## An Example of a Global Model

<table>
<thead>
<tr>
<th>Topic</th>
<th>Translation</th>
<th>Topic</th>
<th>Translation</th>
<th>Topic</th>
<th>Translation</th>
<th>Topic</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>सुनामी, भूकंप, चाइल, पिचिलेमू, गाये, चेतावनी, खबर, शहर</td>
<td>Tsunami, earthquake, Chile, Pichilemu, gone, warning, news, city</td>
<td>विमान, एयर, फ्रांस, जहाज़, ब्राज़ील, ए, 447, गायब, महासागर, फ्रांसीसी</td>
<td>flight, Air, France, Brazil, A, 447, disappear, ocean France</td>
<td>चीन, ओलंपिक, बीजिंग, गोर, समारोह, स्वर्ण, स्टेडियम, खेलों</td>
<td>China, Olympic, Beijing, Gore, function, stadium, games</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
National Institute of Standards and Technology

From Wikipedia, the free encyclopedia

"NIST" redirects here. For other uses, see NIST (disambiguation).

The National Institute of Standards and Technology (NIST), known between 1901 and 1988 as the National Bureau of Standards (NBS), is a measurement standards laboratory which is a non-regulatory agency of the United States Department of Commerce. The institute's official mission is to:[1]

Promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life.

Categories: Standards organizations | National Institute of Standards and Technology | Gaithersburg, Maryland | United States Department of Commerce agencies | Government agencies established in 1901
It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director, Mr. Lopez, stole industrial secrets from the US group and took them with him when he joined VW last year.

This investigation was launched by US President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of Ms. Holland.

Dorothea Holland, until four months ago was the only prosecuting lawyer on the German case.
It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director, Mr. Lopez, stole industrial secrets from the US group and took them with him when he joined VW last year.

This investigation was launched by US President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of Ms. Holland.

Dorothea Holland, until four months ago was the only prosecuting lawyer on the German case.
What helped us generate the Document Level perspective?

The “word level” perspective

It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director Mr. Lopez, stole industrial secrets from the US group and took them with him when he joined VW last year.

This investigation was launched by US President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of Ms. Holland.

Dorothea Holland, until four months ago was the only prosecuting lawyer on the German case.
What if we turn the document off?

- Summarization power of the perspectives

- **US** investigators have asked, but have been so far refused access to, evidence accumulated by **German** prosecutors probing allegations that former **GM** director, Mr. Lopez, stole industrial secrets from the **US** group and took them with him when he joined **VW** last year.

- This investigation was launched by **US** President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of **Ms. Holland** Dorothea Holland, until four months ago was the only prosecuting lawyer on the **German** case.
Assumptions of the Global Models

• Documents are at least tagged from two different perspectives – either implicit or explicit and one perspective affects the other
  – Simplest example of implicit WL tagging – binned positions indicating sections
  – Simplest example of implicit DL tagging – tag cloud

It is believed US investigators have asked for, but have been so far refused access to, evidence accumulated by German prosecutors probing allegations that former GM director, Mr. Lopez, stole industrial secrets from the US group and took them with him when he joined VW last year.

This investigation was launched by US President Bill Clinton and is in principle a far more simple or at least more single-minded pursuit than that of Ms. Holland.

Dorothea Holland, until four months ago was the only prosecuting lawyer on the German case.
### Document Level Perspectives

#### Guided Summarization Track

<table>
<thead>
<tr>
<th>Centers of Attentions (with regard to grammatical or semantic roles)</th>
<th>Menu_foods:ne-&gt;ne, pet:nn-&gt;nn, unit:nn-&gt;nn, Henderson:ne-&gt;ne, wheat:nn-&gt;nn, food:subj-&gt;nn etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 20 (tf-idf)\text{docset} words + Top 5 most frequent non-stopwords in the documents</td>
<td>Menu_Foods, pet, associate, plant, sell, source, FDA, Henderson, agency, shelf, test, unit, Canadian, dog, food etc.</td>
</tr>
</tbody>
</table>

#### Multilingual Track

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 20 (tf-idf)\text{docset} words + Top 5 most frequent non-stopwords in the documents</td>
<td>जांच(investigation), घरों(houses), तलाशी(search), पुलिस(police), स्टेशन(station), किंग्स(King’s), क्रॉस(Cross), हमले(attack) etc.</td>
</tr>
</tbody>
</table>
Word Level Perspectives

• Guided Summarization Track
  – Named Entity classes (Person, Organization, Location, Misc, Date/time/money/number/ordinal/percent)
  – Subjective class e.g. “Of the 10 cats and dogs whose deaths have been linked to pet food that was recalled over the weekend, seven died last month in a taste test conducted by...”

• Multilingual Track
  – {0, 1, 2, 3, 4}: Words annotated by positional bins – document segregated into 5 “sections”
Global(Background) Models

- **METag\(^2\)LDA**: A topic generating all DL tags in a document doesn’t necessarily mean that the same topic generates all words in the document.
- **CorrMETag\(^2\)LDA**: A topic generating *all* DL tags in a document does mean that the same topic generates all words in the document.

The idea was to assign weights to words in sentences from a generative standpoint.
Global (Background) Models

METag\(^2\)LDA

CorrMETag\(^2\)LDA

PER, ORG etc..
Local Models - Guided

- **Guided Summarization Track**
  - Collection of a bag of all nouns (Bag-nn) which are not proper nouns from the Document Level perspective
  - Collection of a bag of all verbs (Bag-vb) which are not stopwords from the Word Level perspective
  - Collection of the dependency parsing (using open-source Stanford CoreNLP parser) outputs for each sentence in the docset

- \[ \text{score}_{\text{global}}(q, \text{sentence}_i) = \sum_{j=1}^{Nq(i)} \sum_{k=1}^{\text{latent topics}} p(w_{q(i)}, z_k) \] where \( \text{sentence}_i \) is within a context that is fit to the model METag^2 LDA

- Finally \( \text{score}_{\text{sentence}_i} \) by greedily optimizing:
  \[ F = \text{score}_{\text{global}+\text{local}}(\text{query}, \text{sentence}_i) - \text{redundancy} (\text{sentence}_i, \text{sentence}_j) - \text{redundancy} (\text{sentence}_i, \text{prev}_{\text{summary}}) \delta \{1, \text{has}_{\text{prev}_{\text{summary}}}\} \]

- Overlapping sentence removal and heuristic sentence pruning afterwards
Local Model - Multilingual

• Multilingual Track
  – Purely based on probabilities!
    • Take a multilingual sentence context \((s \pm 1)\) where the central sentence is at least 10 words long
    • Obtain the log likelihoods of the contexts to the trained corrMETag\(^2\)LDA model with 30 topics
    • Order the sentences in descending order of likelihoods
    • Post-processing only involves keeping sentences within a length threshold, checking for overlaps and removing sentences beginning with a quote
Multilingual Track: Automatic scoring

Our system ID: 7 – Not doing well by just fitting sentence contexts likelihoods
Our system ID: 7 – Doing average by just fitting sentence contexts likelihoods

But scores are stable across most languages
Conclusions

• Experiment with $Sum_{CF}$ as in Nenkova et. al., 2006 over all “important” sentential words not just query words
• Improve and add simple but effective local models that uses closed class words like stopwords
• Verb identification in multilingual documents can help local models