Contextually Mediated Semantic Similarity Graphs for Topic Segmentation

Geetu Ambwani & Tony Davis

StreamSage/Comcast
Outline of talk

- Motivations
- Relevance intervals
- Graphs representing documents
  - Application to segmentation
- Experiments and Evaluation
  - Comparison with other systems
- Conclusions and future work
Topic segmentation

- Topic segmentation defined: dividing a document into topically coherent segments
  - Typically a partition (exhaustive, non-overlapping segments)
  - But could vary (e.g., hierarchical, overlapping, “fuzzy”, etc.)
  - Labeling the segments with good terms is a separate problem

- Advantages of segmenting video (e.g., news broadcasts)
  - Viewers can select only the portions of a program they want to watch
  - They can browse in the order they want
Related Work on Segmentation

- Previous work has used several approaches
  - Discourse features
    - Some signal a topic shift; others a continuation
    - Highly domain-specific
  - Similarity measures between adjacent blocks of text
    - Typical document similarity measures used, as in TextTiling (Hearst, 1994) or Choi’s algorithm (Choi, 2000)
    - Choi measures lexical similarity among neighboring sentences
    - Posit boundaries at points where similarity is low
  - Lexical chains: repeated occurrences of a term (or of closely related terms)
    - Again, posit boundaries where cohesion is low (few lexical chains cross the boundary (e.g., Galley, et al., 2003))
Motivations behind our approach

- Model both the influence of a term beyond the sentence it occurs in and semantic relatedness among terms
  - The range of a term’s influence extends beyond the sentence it occurs in, but how far? (relevance intervals)
  - Semantic relatedness among terms (contextually mediated graphs)
- Apply this model to topic-based segmentation
Relevance Intervals
Relevance Intervals (RIs)

- Each RI is a contiguous segment of audio/video deemed relevant to a term
- Developed originally to improve audio/video search and retrieval
- RI calculation relies on a pointwise mutual information (PMI) model of term co-occurrence (built from 7 years of *New York Times* text, 325M words)
- Previously evaluated on radio news broadcasts, and currently deployed in Comcast video search

\[
PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}
\]

Relevance Intervals (RIs)

- Each RI is a contiguous segment of audio/video deemed relevant to a term
  - RIs are calculated for all content words (after lemmatization) and common multi-word expressions
  - An RI for a term is built outwards, forward and backward from a sentence containing that term, based on:
    - PMI values between pairs of terms across sentences; high PMI values suggest semantic similarity between terms
    - Discourse markers which extend or end an RI
    - Synonym-based query expansion, using information from WordNet
    - Anaphor resolution – roughly based on Kennedy and Boguraev (1996)
    - Nearby RIs for the same term are merged
    - Large-scale vocabulary shifts (as determined by a modified version of Choi (2000) to indicate boundaries)****
Relevance Intervals: an Example

- Index term: **squatter**
  among the sentences containing this term are these two, near each other:

  Paul Bew is professor of Irish politics at Queens University in Belfast. In South Africa the government is struggling to contain a growing demand for land from its black citizens. Authorities have vowed to crack down and arrest **squatters** illegally occupying land near Johannesburg. In a most serious incident today more than 10,000 black South Africans have seized government and privately-owned property. Hundreds were arrested earlier this week and the government hopes to move the rest out in the next two days. NPR’s Kenneth Walker has a report. Thousands of **squatters** in a suburb outside Johannesburg cheer loudly as their leaders deliver angry speeches against whites and landlessness in South Africa. “Must give us a place…”

- We build an RI for **squatter** around each of these sentences…
Index term: **squatter**

among the sentences containing this term are these two, near each other:

```
Paul Bew is professor of Irish politics at Queens University in Belfast.

In South Africa the government is struggling to contain a growing demand for land from its black citizens. Authorities have vowed to crack down and arrest squatters illegally occupying land near Johannesburg.

In a most serious incident today more than 10,000 black South Africans have seized government and privately-owned property. Hundreds were arrested earlier this week and the government hopes to move the rest out in the next two days.

NPR’s Kenneth Walker has a report.
```

Thousands of squatters in a suburb outside Johannesburg cheer loudly as their leaders deliver angry speeches against whites and landlessness in South Africa.

```
“Must give us a place…”
```

We build an RI for **squatter** around each of these sentences…
Relevance Intervals: an Example

- Index term: **squatter**
  among the sentences containing this term are these two, near each other:

  Paul Bew is professor of Irish politics at Queens University in Belfast.

  In South Africa the government is struggling to contain a growing demand for land from its black citizens. Authorities have vowed to crack down and arrest **squatters** illegally occupying land near Johannesburg.

  In a most serious incident today more than 10,000 black South Africans have seized government and privately-owned property. Hundreds were arrested earlier this week and the government hopes to move the rest out in the next two days. NPR’s Kenneth Walker has a report.

  Thousands of **squatters** in a suburb outside Johannesburg cheer loudly as their leaders deliver angry speeches against whites and landlessness in South Africa.

  “Must give us a place…”

  The two intervals for **squatter** are merged, because they are so close
Documents → Graphs → Segmentation
(S_1) Yesterday, I took my dog to the park.
(S_2) While there, I took him off the leash to get some exercise.
(S_3) After 2 minutes, Spot began chasing a squirrel.
(Topic Shift)
(S_4) Then, I needed to go grocery shopping.
(S_5) So I went later that day to the local store.
(S_6) Unfortunately, they were out of cashews.
RIs $\rightarrow$ Nodes

- Construct a graph in which each node represents a term and a sentence, iff the sentence is contained in an RI for that term.
RIs $\rightarrow$ Nodes

- Construct a graph in which each node represents a term and a sentence, iff the sentence is contained in an RI for that term.

Nodes corresponding to these Relevance Intervals

- Sentence 1: dog, park
- Sentence 2: dog, park, leash, exercise
- Sentence 3: dog, park, leash, exercise
Connecting the Nodes ...

All edge strengths between a term and itself are initialized to 1.0.

Sentence 1

Sentence 2

Sentence 3

leash

exercise

leash

exercise
Calculating connection strengths for edges

For edges between different terms, initialize their strengths to normalized PMI values: $s(x, y) = 1 - 1/\exp(PMI(x, y))$
Calculating connection strengths for edges

Add $s('park','leash)s('leash','dog')$ to edge strength between 'park' and 'dog'
Connection strength formula

Connection-strength(A,B) = 2s(A,B) + s(A,X)s(X,B) + s(B,y)s(Y,A)

and in general, for terms a and b in sentences i and i + 1 respectively:

\[ c(a,b) = \sum_{x \in W_i} s(x,a)s(x,b) + \sum_{x \in W_{i+1}} s(y,a)s(y,b) \]
Filtering edges in the graph

- We filter out edges with a connection strength below a set threshold (we’ve tried a couple and usually use 0.5)
Graph Representation of Document

- Let's look at a real example. 1st 8 minutes of an episodes of Bizarre Foods.
- [Bizarre_Foods_With_Andrew_Zimmern-Japan.pdf](Bizarre_Foods_With_Andrew_Zimmern-Japan.pdf)
Segmentation from graphs

- General idea: look for places in the graph where connections are sparse or weak
  - Typically, this will be where relatively few Ris cross a boundary
  - Edges with low connection strengths are unlikely to bear on topical coherence, so it’s best to remove them from the graph

- “Normalized novelty”: on the two sides of a potential boundary, the number of nodes labeled with the same terms, normalized by the total number of terms
Graph representation of documents

Example snippet and graph from t.v. news broadcast

S_190  We’ve got to get this addressed and hold down health care costs.

S_191  Senator ron wyden, the optimist from oregon, we appreciate your time tonight.

S_192  Thank you.

S_193  Coming up, the final day of free health clinic in kansas city, missouri.
Experiments and Evaluation
Evaluation metrics

- How well does the hypothesized set of boundaries match the true (reference) set?
- \( P_k \) (Beeferman, et al. 1997) and WindowDiff (Pevzner & Hearst, 2002)
  - Both compare hypothesis to reference segmentation within a sliding window
  - \( P_k \) is the proportion of windows in which hypothesis and reference disagree on the number of boundaries
  - WindowDiff tallies the difference in the number of boundaries in each window
  - Both commonly used instead of precision and recall, because they take approximate matching into account
  - They have drawbacks of their own, however


Evaluation metrics

- $P_k$ and WindowDiff: sliding window is half the average reference segment size
Evaluation metrics

- One black mark against the hypothesis segmentation, where it differs from the reference (mistakes closer to reference boundaries appear in fewer windows, and are thus penalized less)
# Systems compared

<table>
<thead>
<tr>
<th>Choi</th>
<th>Implementation from MorphAdorner*</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN</td>
<td>Our system, using a single node for each term occurrence (no extension)</td>
</tr>
<tr>
<td>FE</td>
<td>Our system, using an extension of a fixed number of sentences for each term from the sentence it occurs in</td>
</tr>
<tr>
<td>SS</td>
<td>Our system, using Ris without “hard” boundaries determined by the modified Choi algorithm</td>
</tr>
<tr>
<td>SS+C</td>
<td>Our full segmentation system, incorporating “hard” boundaries determined by the modified Choi algorithm</td>
</tr>
</tbody>
</table>

* morphadorner.northwestern.edu/morphadorner/-textsegmenter
Results on pseudodocuments

185 documents each containing 20 Concatenated *New York Times* articles
Number of boundaries not specified to systems

<table>
<thead>
<tr>
<th>system</th>
<th>precision</th>
<th>recall</th>
<th>F</th>
<th>Pk</th>
<th>WindowDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choi</td>
<td>0.404</td>
<td>0.569</td>
<td>0.467</td>
<td>0.338</td>
<td>0.360</td>
</tr>
<tr>
<td>SN</td>
<td>0.096</td>
<td>0.112</td>
<td>0.099</td>
<td>0.570</td>
<td>0.702</td>
</tr>
<tr>
<td>FE</td>
<td>0.265</td>
<td>0.140</td>
<td>0.176</td>
<td>0.478</td>
<td>0.536</td>
</tr>
<tr>
<td>SS</td>
<td>0.566</td>
<td>0.383</td>
<td>0.448</td>
<td>0.292</td>
<td>0.317</td>
</tr>
<tr>
<td>SS+C</td>
<td>0.578</td>
<td>0.535</td>
<td>0.537</td>
<td>0.262</td>
<td>0.283</td>
</tr>
</tbody>
</table>
Results on TV shows

- Data: Closed captions for 13 tv shows (News, talk shows, documentaries, lifestyle shows)
- 5 annotators manually marked up major and minor boundaries, using 1-5 rating scale
- As expected, IAA is low, so we create a reference annotation

```
annotator A
annotator B
annotator C

majority of annotators mark a boundary within a window

“reference” annotation places a boundary according to majority vote
```
TV show closed-captions: inter-annotator agreement on segmentation

- $P_k$ values between pairs of annotators: all boundaries and *major boundaries*
- Note that matrix is asymmetrical

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>0.36</td>
<td>0.30</td>
<td>0.27</td>
<td>0.42</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.48</td>
<td>0.45</td>
<td>0.44</td>
<td>0.67</td>
<td>0.38</td>
</tr>
<tr>
<td>B</td>
<td>0.29</td>
<td></td>
<td>0.29</td>
<td>0.27</td>
<td>0.33</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td></td>
<td>0.32</td>
<td>0.33</td>
<td>0.55</td>
<td>0.25</td>
</tr>
<tr>
<td>C</td>
<td>0.57</td>
<td>0.60</td>
<td></td>
<td>0.41</td>
<td>0.67</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>0.48</td>
<td>0.44</td>
<td></td>
<td>0.20</td>
<td>0.61</td>
<td>0.18</td>
</tr>
<tr>
<td>D</td>
<td>0.36</td>
<td>0.41</td>
<td>0.27</td>
<td></td>
<td>0.53</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>0.46</td>
<td>0.46</td>
<td>0.20</td>
<td></td>
<td>0.63</td>
<td>0.26</td>
</tr>
<tr>
<td>E</td>
<td>0.33</td>
<td>0.31</td>
<td>0.33</td>
<td>0.32</td>
<td></td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.34</td>
<td>0.30</td>
<td>0.31</td>
<td></td>
<td>0.27</td>
</tr>
<tr>
<td>Ref</td>
<td>0.25</td>
<td>0.32</td>
<td>0.24</td>
<td>0.21</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.39</td>
<td>0.35</td>
<td>0.17</td>
<td>0.22</td>
<td>0.58</td>
<td></td>
</tr>
</tbody>
</table>
TV show closed-captions: segmentation

- Accuracy is low, which is unsurprising given the low IAA

<table>
<thead>
<tr>
<th>system</th>
<th>precision</th>
<th>recall</th>
<th>F</th>
<th>Pk</th>
<th>WindowDiff</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All topic boundaries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choi</td>
<td>0.197</td>
<td>0.186</td>
<td>0.184</td>
<td>0.476</td>
<td>0.507</td>
</tr>
<tr>
<td>SS+C</td>
<td>0.315</td>
<td>0.208</td>
<td>0.240</td>
<td>0.421</td>
<td>0.462</td>
</tr>
<tr>
<td>Major topic boundaries only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choi</td>
<td>0.170</td>
<td>0.296</td>
<td>0.201</td>
<td>0.637</td>
<td>0.812</td>
</tr>
<tr>
<td>SS+C</td>
<td>0.271</td>
<td>0.316</td>
<td>0.271</td>
<td>0.463</td>
<td>0.621</td>
</tr>
</tbody>
</table>
Conclusions and future work
Conclusions and future work

Conclusions

- Graphs constructed from RIs do seem to help segmentation
- Semantic relatedness with reinforcement from neighboring terms
- Works decently on “noisy” material, such as TV shows
- Doesn’t require any training; however, there are lots of parameters to play with (and we have started exploring training to optimize them)

Future work

- Several ways to segment a graph: try community detection or learn boundary detection through various graph features
- Try to use graphs for more complex segmentation tasks, such as hierarchical segmentation; community structure in a graph might reflect hierarchical organization of discourse
- Try to find the most “central” terms in a subgraph, to use as segment labels
We gratefully acknowledge…

Gene Chipman
Oliver Jojic
Akash Nagle
Robert Rubinoff
Hassan Sayyadi

Thank you! Questions?