Graph-based Neural Multi-Doc Summarization

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, Dragomir Radev
Text Summarization

The process of **shortening** a text to create a summary with the **major points**

- Single-doc summarization
- **Multi-doc summarization (MDS):**
  - A cluster of documents about the same topic

**Methods**

- **Extractive:**
  - Extract salient sentences (phrases) from the original text
- **Abstractive:**
  - Generate a summary from scratch, involving paraphrasing
Why Multi-Doc Summarization (MDS)?

Often times, we want a summary for a whole topic, rather than one document.

- E.g. different news articles about the same event

- More challenging, as we need to think about the relationships between documents
Recent Models for Single-Doc Summarization

RNN encoder-decoder model
- E.g. Cheng and Lapata (2016),

See et al. (2017) - abstractive
Can we extend this to Multi-Docs?

Several issues:

- **RNN encoder**
  - A simple RNN cannot capture the relationships between different documents
  - Christensen et al. (2013) demonstrates the importance of considering sentence relations among multiple documents

- **RNN decoder**
  - Requires large training data, which MDS datasets are lacking in
  - Instead, we break down the summarization task into sentence salience estimation and sentence selection
Graph Representation of A Document Cluster

Sentence relation graph

- node = sentence
- edge / weight
  = relation between two sentences
Graph Representation Methods

- **Cosine Similarity Graph:**
  - LexRank (Erkan and Radev, 2004)
  - Edge weight: tf-idf cosine similarity between sentences

- **Approximate Discourse Graph (ADG):**
  - Christensen et al. (2013)

- **Personalized Discourse Graph (PDG):**
  - Our modification of ADG
Approximate Discourse Graph (ADG)

- Christensen et al. (2013)

- Construct edges between sentences by counting **discourse relation indicators:**
  - deverbal noun references, event and entity continuations, discourse markers, co-referent mentions, etc.

- Edge weight is discrete & not diverse
  - Often 0, 1, or 2
Personalized Discourse Graph (PDG)

- Modification of ADG: **Diversify edge weights** via sentence personalization

- Sentence personalization score:
  - Consider **weighting for each sentence**
  - Computed by simple macro-level features (sentence position, length, etc.)
    (Christensen et al. 2013)

- Edge weight is transformed via

\[
\ell_{PDG}(u, v) = \frac{\ell_{ADG}(u, v) s(u)}{\sum_{u' \in V} \ell_{ADG}(u', v) s(u')}
\]

where \( s(\cdot) \) is the personalization scores
Our Model

- **Neural Net** for powerful representation learning
- **Graph** for capturing the sentence relations in a document cluster
Our Model

Sentence Relation Graph

Cluster

doc1
- d1s1
- d1s2

doc2
- d2s1
- d2s2

Sentence Embedding

GRU sent

h_0 \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow h_4

w_1 \rightarrow w_2 \rightarrow w_3
Our Model - GCN

Graph Convolutional Networks (GCN)  
(Kipf and Welling, 2017)

\[ H^{(l+1)} = \sigma \left( \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right) \]

A: adjacency matrix  
H: node features  
W: learnable weight
Our Model - GCN

Graph Convolutional Networks (GCN)
(Kipf and Welling, 2017)
Our Model - Salience Estimation
Training & Testing

**Training:**
- Obtain target salience score for each sentence:
  - Average of the ROUGE 1 & 2 scores
- Minimize the loss between the *estimated* scores and the *target* scores

**Testing:**
- Select sentences with high predicted scores in a *greedy manner*
- Avoid redundancy during selection:
  - Add the sentence if its cosine similarity to current summary is < 0.5
- (Hong & Nenkova, 2014)
## Experiments - Data set


<table>
<thead>
<tr>
<th></th>
<th>DUC’01</th>
<th>DUC’02</th>
<th>DUC’03</th>
<th>DUC’04</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Clusters</td>
<td>30</td>
<td>59</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td># of Documents</td>
<td>309</td>
<td>567</td>
<td>298</td>
<td>500</td>
</tr>
<tr>
<td># of Sentences</td>
<td>24498</td>
<td>16090</td>
<td>7721</td>
<td>13270</td>
</tr>
<tr>
<td>Vocabulary Size</td>
<td>28188</td>
<td>22174</td>
<td>13248</td>
<td>18036</td>
</tr>
<tr>
<td>Summary Length</td>
<td>100 words</td>
<td>100 words</td>
<td>100 words</td>
<td>665 Bytes</td>
</tr>
</tbody>
</table>
Experiments - Evaluation

Metric: ROUGE-1, 2 (Lin, 2004)

Model comparison

Baseline: vanilla GRU (NO graph)

Graph-based neural:
- GRU + GCN (cosine similarity graph)
- GRU + GCN (ADG)
- GRU + GCN (PDG)
Research Questions

- Does the GCN model with sentence relation graphs improve upon the vanilla RNN model?

- How does our model compare with traditional graph-based summarizers?

- Comparison of different sentence relation graphs:
  - Cosine Similarity Graph
  - ADG
  - PDG
## Results

### Our models

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRU only (No Graph)</td>
<td>36.04</td>
<td>8.47</td>
</tr>
<tr>
<td>GCN: Similarity Graph</td>
<td>37.33</td>
<td>8.78</td>
</tr>
<tr>
<td>GCN: ADG</td>
<td>37.41</td>
<td>8.97</td>
</tr>
<tr>
<td>GCN: PDG</td>
<td><strong>38.23</strong></td>
<td><strong>9.48</strong></td>
</tr>
</tbody>
</table>

### Other baseline & SOTA multi-doc summarizers

<table>
<thead>
<tr>
<th>Model</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid (Radev et al., 2007)</td>
<td>36.41</td>
<td>7.97</td>
</tr>
<tr>
<td>CLASSY 04 (Conroy et al., 2007)</td>
<td>37.62</td>
<td>8.96</td>
</tr>
<tr>
<td>LexRank (Erkan and Radev, 2004)</td>
<td>35.95</td>
<td>7.47</td>
</tr>
<tr>
<td>G-Flow (Christensen et al., 2013)</td>
<td>35.30</td>
<td>8.27</td>
</tr>
<tr>
<td>SVR (Li et al., 2007)</td>
<td>36.18</td>
<td>9.34</td>
</tr>
<tr>
<td>RegSum (Hong &amp; Nenkova, 2014)</td>
<td><strong>38.57</strong></td>
<td><strong>9.75</strong></td>
</tr>
</tbody>
</table>
Analysis - Learning Curve

PDG
- converges the fastest
- shows the best generalization
**Analysis - Graph Characteristics**

- **PDG**: stronger correlation, thanks to the improvement in edge weight
- **ADG**: discrete weight
- **Similarity Graph**: weaker correlation

---

**x axis = salience, y axis = node degree**
Conclusion & Contribution

Our GCN model performed ...

1. better than vanilla GRU
   => Graph captures *sentence relations across documents*

2. better than traditional graph-based summarizers
   => Powerful representation learning by neural net

Gateway to incorporating graph-based techniques into neural summarization
Conclusion & Contribution

Our **Personalized Discourse Graph (PDG)** provides the most informative guide to salience estimation, among the three graph representation methods.

=> Advantage of *discourse relations* over simple similarity
   Efficacy of adjusting edge weights via *sentence personalization*
Thank you!

michihiro.yasunaga@yale.edu