Cascaded Attention based Unsupervised Information Distillation for Compressive Summarization

Piki Li, Wai Lam, Lidong Bing, Weiwei Guo, Hang Li
The Chinese University of Hong Kong; AI Lab, Tencent Inc.; LinkedIn; Noah’s Ark Lab, Huawei Technologies
EMNLP 2017

Presenter: Tzu-En Liu
Outline

- Introduction
- Framework Description
- Experimental Setup
- Results and Discussion
- Conclusion
INTRODUCTION
Multi-Document Summarization

- The goal is to automatically produce a succinct summary, preserving the most important information of a set of documents describing a topic.
Motivation

- When people read, they will remember and forget part of the content.
- The important information usually attracts more attention since it may repeatedly appears in some documents, or be positioned in the beginning paragraph.
Ideas

- Distill salient information from the input documents on an unsupervised data reconstruction manner.
- The word salience is fed into a coarse-grained sentence compression component.
- The attention weights are integrated into a phrase-based optimization framework for compressive summary generation.
Contributions

- Propose a cascaded attention model
- The attention weights are learned automatically by an unsupervised data reconstruction
- Achieve better performance than the state-of-the-art models
FRAMEWORK DESCRIPTION
Overview
Employ the bag-of-words (BOW) representation as the initial semantic representation for sentences.

\[
c_g = h_m^e \\
h_t^e = f(h_{t-1}^e, h_t^v) \\
h_j^v = \tanh(W_{xh}^v x_j + b_h^v)
\]
Employ the bag-of-words (BOW) representation as the initial semantic representation for sentences.

\[ c_g = h_m^e \]
\[ h_t^e = f(h_{t-1}^e, h_t^v) \]
\[ h_j^v = \tanh(W_{xh}^v x_j + b_h^v) \]
Employ the bag-of-words (BOW) representation as the initial semantic representation for sentences.

\[
c_g = h_m^e
\]

\[
h_t^e = f(h_{t-1}^e, h_t^v)
\]

\[
h_j^v = \tanh(W_{xh}^v x_j + b_h^v)
\]
• Employ the bag-of-words (BOW) representation as the initial semantic representation for sentences

\[ c_g = h_m^e \]

\[ h_t^e = f(h_{t-1}^e, h_t^v) \]

\[ h_j^v = \tanh(W_{xh}^v x_j + b_h^v) \]
Recaller

- Reverse of the reader stage

\[
\begin{align*}
    s_t &= \sigma(W_{hs}h_t^o + b_s) \\
    h_t^o &= \tanh(W_{hh}h_t^d + b_h^o) \\
    h_t^d &= f(h_{t-1}^d, c_g)
\end{align*}
\]
Recaller

- Reverse of the reader stage

\[ s_t = \sigma(W_{hs}h_t^o + b_s) \quad S \]

\[ h_t^o = \tanh(W_{hh}h_t^d + b_h^o) \quad H^o \]

\[ h_t^d = f(h_{t-1}^d, c_g) \quad \text{Dec} \]
Recaller

- Reverse of the reader stage

\[
\begin{align*}
    s_t &= \sigma(W_{hs}h^o_t + b_s) \\
    h^o_t &= \tanh(W^o_{hh}h^d_t + b^o_h) \\
    h^d_t &= f(h^d_{t-1}, c_g)
\end{align*}
\]
Cascaded Attention Modeling

- Add attention mechanism in the hidden layer and the output layer
- Capture the salience of sentences from two different and complementary vector spaces
  - The embedding space - provides better generalization
  - The BOW vector space - captures more nuanced and subtle difference
Cascaded Attention Modeling

- The embedding space

\[
\text{score}(h_t, h_s) = v^T \tanh(W[h_t; h_s])
\]

\[
a_{t,i}^h = \frac{\exp(\text{score}(h_t^o, h_i^v))}{\sum_{i'} \exp(\text{score}(h_t^o, h_{i'}^v))}
\]

\[
c_t^h = \sum_{i'} a_{t,i}^h h_{i'}^v
\]

\[
\tilde{h}_t^o = h_t^o
\]

\[
h_t^o = \tanh(W_{ch}^a c_t^h + W_{hh}^a \tilde{h}_t^o)
\]
Cascaded Attention Modeling

- The BOW vector space

\[
\bar{a}_{t,i}^o = \frac{\exp(score(s_t, x_i))}{\sum_{i'} \exp(score(s_t, x_{i'}))}
\]

\[
a_{t,i}^o = \lambda_a \bar{a}_{t,i}^o + (1 - \lambda_a) a_{t,i}^h
\]

\[
c_t^o = \sum_{i'} a_{t,i'} x_{i'}
\]

\[
\tilde{s}_t = s_t
\]

\[
s_t = \lambda_c c_t^o + (1 - \lambda_c) \tilde{s}_t
\]
Unsupervised Learning

- Word salience score is the magnitude vector computed from the columns in S
- Sentence salience score is the magnitude vector computed from the columns in A°

\[
\min_{\Theta} \frac{1}{2m} \sum_{i=1}^{m} \|x_i - \sum_{j=1}^{n} s_j a_{j,i}^o\|_2^2 + \lambda_s \|S\|_1
\]
Coarse-Grained Sentence Compression

- Jointly considers word salience obtained from the neural attention model and linguistically-motivated rules
- The linguistically-motivated rules are designed based on the observed obvious evidence for uncritical information from the word level to the clause level
- Information with smaller salience score will be removed
Phrase-Based Optimization For Summary Construction

- Extract the noun-phrases (NPs) and verb-phrases (VPs)

\[ S_i = \left\{ \sum_{t \in P_i} \frac{tf(t)}{\sum_{t \in \text{Topic}} tf(t)} \right\} \times a_i \]

the frequency of the concept \( t \) (unigram/bigram)
Phrase-Based Optimization For Summary Construction

- Extract the noun-phrases (NPs) and verb-phrases (VPs)

\[
S_i = \left\{ \frac{\sum_{t \in P_i} tf(t)}{\sum_{t \in \text{Topic}} tf(t)} \right\} \times a_i
\]

salience of the sentence containing \( P_i \)
Phrase-Based Optimization For Summary Construction

- Extract the noun-phrases (NPs) and verb-phrases (VPs)

\[ S_i = \left\{ \frac{\sum \text{tf}(t)}{\sum \text{tf}(t)} \right\} \times a_i \]

- The overall objective function

\[
\max \left\{ \sum_i \alpha_i S_i - \sum_{i<j} \alpha_{ij} (S_i + S_j) R_{ij} \right\}
\]

selection indicator
Phrase-Based Optimization For Summary Construction

- Extract the noun-phrases (NPs) and verb-phrases (VPs)

\[ S_i = \left\{ \sum_{t \in P_i} tf(t) / \sum_{t \in Topic} tf(t) \right\} \times a_i \]

- The overall objective function

\[ \max \left\{ \sum_i \alpha_i S_i - \sum_{i<j} \alpha_{ij} (S_i + S_j) R_{ij} \right\} \]

co-occurrence indicator
Phrase-Based Optimization For Summary Construction

- Extract the noun-phrases (NPs) and verb-phrases (VPs)

\[
S_i = \left\{ \sum_{t \in P_i} tf(t) / \sum_{t \in \text{Topic}} tf(t) \right\} \times a_i
\]

- The overall objective function

\[
\max\left\{ \sum_i \alpha_i S_i - \sum_{i<j} \alpha_{ij} (S_i + S_j) R_{ij} \right\}
\]

Similarity of a pair of phrases (Jaccard Index based method)
Phrase-Based Optimization For Summary Construction

- Extract the noun-phrases (NPs) and verb-phrases (VPs)

\[ S_i = \left\{ \frac{\sum_{t \in P_i} tf(t)}{\sum_{t \in \text{Topic}} tf(t)} \right\} \times a_i \]

- The overall objective function

\[ \max \{ \sum_i \alpha_i S_i \} - \sum_{i<j} \alpha_{ij} (S_i + S_j) R_{ij} \]

maximize the salience score of the selected phrases  
penalize the selection of similar phrase pairs
Phrase-Based Optimization For Summary Construction

- Extract the noun-phrases (NPs) and verb-phrases (VPs)

\[ S_i = \left\{ \sum_{t \in P_i} tf(t) / \sum_{t \in \text{Topic}} tf(t) \right\} \times a_i \]

- The overall objective function

\[ \max \left\{ \sum_i \alpha_i S_i - \sum_{i<j} \alpha_{ij} (S_i + S_j) R_{ij} \right\} \]

\[ \forall P_i \in x_k, \alpha_i \leq \beta_k \land \sum_i \alpha_i \geq \beta_k \]

(selection indicator of the sentence \(x_k\))
Phrase-Based Optimization For Summary Construction

- Post-processing
  - sentences in summary are ordered according to their natural order if they come from the same document
  - otherwise, they are ordered according to the timestamps of the corresponding documents
EXPERIMENTAL SETUP
Dataset

- DUC 2006 (50 topics, 25 documents, 4 summaries)
- DUC 2007 (45 topics, 25 documents, 4 summaries)
- TAC 2011 (44 topics, 10 documents, 4 summaries)
- TAC 2010 (used as the parameter tuning)
RESULTS AND DISCUSSION
Effect Of Existing Salience Models & Different Attention Architectures

Table 1: Comparisons on TAC 2010

<table>
<thead>
<tr>
<th>System</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW</td>
<td>0.353</td>
<td>0.092</td>
<td>0.123</td>
</tr>
<tr>
<td>SC</td>
<td>0.346</td>
<td>0.083</td>
<td>0.116</td>
</tr>
<tr>
<td>AttenC-tensor-gru</td>
<td>0.339</td>
<td>0.078</td>
<td>0.115</td>
</tr>
<tr>
<td>AttenC-concat-gru</td>
<td>0.353</td>
<td>0.089</td>
<td>0.121</td>
</tr>
<tr>
<td>AttenC-dot-lstm</td>
<td>0.352</td>
<td>0.089</td>
<td>0.121</td>
</tr>
<tr>
<td>AttenH-dot-gru</td>
<td>0.348</td>
<td>0.086</td>
<td>0.119</td>
</tr>
<tr>
<td>AttenO-dot-gru</td>
<td>0.348</td>
<td>0.085</td>
<td>0.118</td>
</tr>
<tr>
<td>AttenC-dot-gru (w/o coarse-comp)</td>
<td><strong>0.359</strong></td>
<td><strong>0.092</strong></td>
<td><strong>0.124</strong></td>
</tr>
<tr>
<td>(w/o coarse-comp)</td>
<td>0.351</td>
<td>0.089</td>
<td>0.122</td>
</tr>
</tbody>
</table>
# Main Results Of Compressive MDS

## Table 2: Results on DUC 2006.

<table>
<thead>
<tr>
<th>System</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.280</td>
<td>0.046</td>
<td>0.088</td>
</tr>
<tr>
<td>Lead</td>
<td>0.308</td>
<td>0.048</td>
<td>0.087</td>
</tr>
<tr>
<td>LexRank</td>
<td>0.360</td>
<td>0.062</td>
<td>0.118</td>
</tr>
<tr>
<td>TextRank</td>
<td>0.373</td>
<td>0.066</td>
<td>0.125</td>
</tr>
<tr>
<td>MDS-Sparse</td>
<td>0.340</td>
<td>0.052</td>
<td>0.107</td>
</tr>
<tr>
<td>DSDR</td>
<td>0.377</td>
<td>0.073</td>
<td>0.117</td>
</tr>
<tr>
<td>RA-MDS</td>
<td>0.391</td>
<td>0.081</td>
<td>0.136</td>
</tr>
<tr>
<td>ABS-Phrase</td>
<td>0.392</td>
<td>0.082</td>
<td>0.137</td>
</tr>
<tr>
<td>C-Attention</td>
<td>0.393*</td>
<td>0.087*</td>
<td>0.141*</td>
</tr>
</tbody>
</table>

## Table 3: Results on DUC 2007.

<table>
<thead>
<tr>
<th>System</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.302</td>
<td>0.046</td>
<td>0.088</td>
</tr>
<tr>
<td>Lead</td>
<td>0.312</td>
<td>0.058</td>
<td>0.102</td>
</tr>
<tr>
<td>LexRank</td>
<td>0.378</td>
<td>0.075</td>
<td>0.130</td>
</tr>
<tr>
<td>TextRank</td>
<td>0.403</td>
<td>0.083</td>
<td>0.144</td>
</tr>
<tr>
<td>MDS-Sparse</td>
<td>0.353</td>
<td>0.055</td>
<td>0.112</td>
</tr>
<tr>
<td>DSDR</td>
<td>0.398</td>
<td>0.087</td>
<td>0.137</td>
</tr>
<tr>
<td>RA-MDS</td>
<td>0.408</td>
<td>0.097</td>
<td>0.150</td>
</tr>
<tr>
<td>ABS-Phrase</td>
<td>0.419</td>
<td>0.103</td>
<td>0.156</td>
</tr>
<tr>
<td>C-Attention</td>
<td>0.423*</td>
<td>0.107*</td>
<td>0.161*</td>
</tr>
</tbody>
</table>

## Table 4: Results on TAC 2011.

<table>
<thead>
<tr>
<th>System</th>
<th>R-1</th>
<th>R-2</th>
<th>R-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.303</td>
<td>0.045</td>
<td>0.090</td>
</tr>
<tr>
<td>Lead</td>
<td>0.315</td>
<td>0.071</td>
<td>0.103</td>
</tr>
<tr>
<td>LexRank</td>
<td>0.313</td>
<td>0.060</td>
<td>0.102</td>
</tr>
<tr>
<td>TextRank</td>
<td>0.332</td>
<td>0.064</td>
<td>0.107</td>
</tr>
<tr>
<td>PKUTM</td>
<td>0.396</td>
<td>0.113</td>
<td>0.148</td>
</tr>
<tr>
<td>ABS-Phrase</td>
<td>0.393</td>
<td>0.117</td>
<td>0.148</td>
</tr>
<tr>
<td>RA-MDS</td>
<td>0.400</td>
<td>0.117</td>
<td>0.151</td>
</tr>
<tr>
<td>C-Attention</td>
<td>0.400*</td>
<td>0.121*</td>
<td>0.153*</td>
</tr>
</tbody>
</table>

* Statistical significance tests show that our method is better.
Linguistic Quality Evaluation

- Grammaticality (Q1), Non-Redundancy (Q2), Referential Clarity (Q3), Focus (Q4), Coherence (Q5)

Table 5: Evaluation of linguistic quality.

<table>
<thead>
<tr>
<th>System</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS-Phrase</td>
<td>3.75</td>
<td>3.38</td>
<td>3.75</td>
<td>3.35</td>
<td>3.12</td>
<td>3.47</td>
</tr>
<tr>
<td>PKUTM</td>
<td>4.13</td>
<td>3.45</td>
<td>3.83</td>
<td>3.33</td>
<td>2.92</td>
<td>3.53</td>
</tr>
<tr>
<td>Ours</td>
<td>3.96</td>
<td>3.50</td>
<td>3.79</td>
<td>3.50</td>
<td>3.25</td>
<td>3.60</td>
</tr>
</tbody>
</table>
Distilled Word Salience

- Topic: “Finland Shooting”, “Heart Disease”, “HIV infection Africa”

Table 6: Top-10 terms extracted from each topic according to the word salience

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>school</td>
<td>heart</td>
<td>HIV</td>
</tr>
<tr>
<td>shooting</td>
<td>disease</td>
<td>Africa</td>
</tr>
<tr>
<td>Auvinen</td>
<td>study</td>
<td>circumcision</td>
</tr>
<tr>
<td>Finland</td>
<td>risk</td>
<td>study</td>
</tr>
<tr>
<td>police</td>
<td>test</td>
<td>infection</td>
</tr>
<tr>
<td>video</td>
<td>blood</td>
<td>trial</td>
</tr>
<tr>
<td>Wednesday</td>
<td>red</td>
<td>woman</td>
</tr>
<tr>
<td>gunman</td>
<td>telomere</td>
<td>drug</td>
</tr>
<tr>
<td>post</td>
<td>level</td>
<td>health</td>
</tr>
</tbody>
</table>
Figure 2: Visualization for sentence attention.
Table 7: The summary of the topic “Hawkins Robert Van Maur”.

S1: The young gunman who opened fire at a mall busy with holiday shoppers appeared to choose his victims at random, according to police[... but a note he left behind hinted at a troubled life].

S2: The teenage gunman who went on a shooting rampage in a department store, [killing eight people,] may have smuggled an assault rifle into the mall underneath clothing[... police said Thursday].

S3: [But] police said it was Hawkins who went into an Omaha shopping mall on Wednesday and began a shooting rampage that killed eight people.

S4: Mall security officers noticed Hawkins briefly enter the Von Maur department store at Omaha’s Westroads Mall earlier Wednesday[... he said].
CONCLUSIONS
Conclusion

- Propose a cascaded neural attention based unsupervised salience estimation method for compressive multi-document summarization
- Investigate the performance of combining different attention architectures and cascaded structures
- Achieves good performance compared with the state-of-the-art methods