Learning Unified Multi-Document Summarization From Collaborative Journalism

Master’s Thesis by Yasar Naci Gündüz

First Referee : Prof.Dr.Benno Stein
Second Referee : Prof.Dr.Andreas Jakoby
INTRODUCTION: New age, new habits

THEN

NOW
INTRODUCTION: New age, new habits

THEN

NOW

ALL THIS TECHNOLOGY IS MAKING US ANTI-SOCIAL
Introduction : How about journalism?

Several research reported:

- Reading attention span is getting shorter
- Young generation is the least informed...
- ...and more interested in social media
Introduction : How about journalism?

Several research reported:

- Reading attention span is getting shorter
- Young generation is the least informed...
- ...and more interested in social media

Information Pollution:

- Reliable sources are more important than ever
Introduction: Our proposal

Make the content:

- Less time consuming
- Yet still adequately informing

Solution: Automatic Summarization
Introduction: Automatic Summarization for Journalism

“Journalism is the activity of gathering, assessing, creating, and presenting news and information.”

American Press Institute
Introduction: Automatic Summarization for Journalism

“Journalism is the activity of gathering, assessing, creating, and presenting news and information.”

- Whole
- Extensive
- Unbiased
Introduction : Automatic Summarization for Journalism

“Journalism is the activity of gathering, assessing, creating, and presenting news and information.”

- Whole
- Extensive
- Unbiased

Solution: Multi-document Summarization
Introduction: Automatic Summarization for Journalism

“Journalism is the activity of gathering, assessing, creating, and presenting news and information.”

- Extractive and Abstractive
Introduction: Automatic Summarization for Journalism

“Journalism is the activity of gathering, assessing, creating, and presenting news and information.”

- Extractive and Abstractive
- Neural Abstractive Summarization
  - Methods are generally for Single-Document
Introduction : Automatic Summarization for Journalism

“Journalism is the activity of gathering, assessing, creating, and presenting news and information.”

- Extractive and Abstractive
- Neural Abstractive Summarization
  - Methods are generally for Single-Document
- Unified Model : Extractive + Abstractive
  - Content Selection
  - Multi-Document -> Single Document
Dataset

Unified Summarization Pipeline

Experiments & Evaluation
Dataset: What do we need?

Neural Abstractive:

- Typically needs a dataset of thousands of documents
- i.e. CNN/Dailymail > 90k/197k (single-document dataset)
## Dataset: What do we have?

- Multi-Document datasets are typically small
- One of the most well-known does not contain more than 60 cluster and 600 documents

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Cluster/Sample Summaries</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUC 2001</td>
<td>30</td>
<td>309</td>
</tr>
<tr>
<td>DUC 2002</td>
<td>59</td>
<td>567</td>
</tr>
<tr>
<td>DUC 2004</td>
<td>50</td>
<td>500</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>139</strong></td>
<td><strong>1,376</strong></td>
</tr>
</tbody>
</table>
Dataset: Solution

- We created Webis-wikinews-corpus
- One of the first of its kind...
  - Large-scale
  - Multi-document
  - For the news domain
Dataset: Source

  - Unbiased
  - Open-source
  - Up-to-date
  - Clustered news from reliable sources
Dataset: Construction

Extract the useful information from Dump File:

- Article, source links, auxiliary information
- Only the pages with news sources for the Wikipedia
Dataset: Construction

Retrieval:
# Dataset: Size & Folder Structure

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Cluster/Sample Summaries</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikinews</td>
<td>9,514</td>
<td>21,314</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>2,174</td>
<td>17,807</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>11,688</strong></td>
<td><strong>39,121</strong></td>
</tr>
</tbody>
</table>
Unified Summarization Pipeline
Unified Summarization

- Extractive Summarization: Wikisummarizer

- Abstractive Summarization: Pointer-Generator Network [See et al., 2017]
Unified Summarization

- **Extractive Summarization: Wikisummarizer**
  - A Google Brain project [Liu et al., 2018]: Extraction from similar source (Wikipedia)

- **Abstractive Summarization: Pointer-Generator Network** [See et al., 2017]
Unified Summarization

- Extractive Summarization: Wikisummarizer
  - A Google Brain project [Liu et al., 2018]: Extraction from similar source (Wikipedia)
  - CST: Filter out the duplication [Radev and Zhang, 2004]
- Abstractive Summarization: Pointer-Generator Network [See et al., 2017]
Unified Summarization

- **Extractive Summarization: Wikisummarizer**
  - A Google Brain project [Liu et al., 2018]: Extraction from similar source (Wikipedia)
  - CST: Filter out the duplication [Radev and Zhang, 2004]
- **Abstractive Summarization: Pointer-Generator Network** [See et al., 2017]
Unified Summarization

- **Extractive Summarization: Wikisummarizer**
  - A Google Brain project [Liu et al., 2018]: Extraction from similar source (Wikipedia)
  - CST: Filter out the duplication [Radev and Zhang, 2004]

- **Abstractive Summarization: Pointer-Generator Network** [See et al., 2017]
  - Solves the problems of earlier approaches such as repetitiveness, senseless sentences and inaccurate facts
Experiments & Evaluation

Training Models:
- double-abstractive
- ea-full-target
- ea-short-target

Evaluation Methods:
- Automatic Evaluation
- Manual Evaluation
- Observations
Experiments and Evaluation: Training Models

- Double-abstractive
- Extractive + Abstractive Full Target
- Extractive + Abstractive Short Target
Experiments and Evaluation: Training Models

Double-abstractive
- Trivial method
- To examine the unified model
Unified Models: Extractive + Abstractive
- ea-full-target - Target document size: Full size
- ea-short-target - Target document size: 3 sentences
- To examine the effects of different ratio between input and target
Introduction: Automatic Summarization for Journalism

“Journalism is the activity of gathering, assessing, creating, and presenting news and information.”
“Journalism is the activity of gathering, assessing, creating, and presenting news and information.”

- Aspects:
  - Content
  - Readability
Experiments and Evaluation: Aspects

- Aspects:
  - Content
    - Automatic > a state-of-the-art method exist
  - Readability
Experiments and Evaluation: ROUGE

Computer Generated Summary: the cat was found under the bed

Ground-truth Summary: the cat was under the bed

\[
\text{Recall} = \frac{\text{number of overlapping text units}}{\text{total words in ground truth}} = \frac{6}{6} = 1
\]
Experiments and Evaluation: ROUGE

Computer Generated Summary: the cat was found under the bed

Ground-truth Summary: the cat was under the bed
Experiments and Evaluation: ROUGE

Computer Generated Summary: the cat was found under the bed

Ground-truth Summary: the cat was under the bed

\[ \text{Precision} = \frac{\text{number of overlapping text units}}{\text{total words in generated text}} = \frac{6}{7} = 0.86 \]
Experiments and Evaluation: ROUGE

Computer Generated Summary: the cat was found under the bed

Ground-truth Summary: the cat was under the bed

\[ F - Measure = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \]
Experiments and Evaluation: ROUGE

- ROUGE-N(ROUGE-1) : Overlapping n-grams > Word wise similarity
- ROUGE-L : Longest Common Subsequence > Sequence wise similarity
Experiments and Evaluation: Results

- Aspects:
  - Content:
    - Automatic > a state-of-the-art method exist

<table>
<thead>
<tr>
<th>ROUGE</th>
<th>double-abstractive</th>
<th>ea-full-target</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1</td>
<td>0.23</td>
<td>0.29</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.16</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Experiments and Evaluation: Results

- Aspects:
  - Content
    - Automatic > a state-of-the-art method exist

<table>
<thead>
<tr>
<th></th>
<th>double-abstractive</th>
<th>ea-full-target</th>
<th>ea-short-target</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1</td>
<td>0.23</td>
<td>0.29</td>
<td>0.54</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.16</td>
<td>0.21</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Experiments and Evaluation: Aspects

- Aspects:
  - **Content**
    - Automatic > a state-of-the-art method exist
  - **Readability**
Experiments and Evaluation: ROUGE for readability?

Computer Generated Summary: was the found under the cat

Ground-truth Summary: the cat was found under the bed

1 ROUGE-1 Average_R: 0.83333
1 ROUGE-1 Average_P: 0.83333
1 ROUGE-1 Average_F: 0.83333

1 ROUGE-L Average_R: 0.50000
1 ROUGE-L Average_P: 0.50000
1 ROUGE-L Average_F: 0.50000
Experiments and Evaluation: ROUGE for readability?

Computer Generated Summary: **was the found under the cat**

Ground-truth Summary: **the cat was found under the bed**

1 ROUGE-1 Average_R: 0.83333
1 ROUGE-1 Average_P: 0.83333
1 ROUGE-1 Average_F: 0.83333

1 ROUGE-L Average_R: 0.50000
1 ROUGE-L Average_P: 0.50000
1 ROUGE-L Average_F: 0.50000

Computer Generated Summary: **he found no lights on**

Ground-truth Summary: all of the lamps were off already when **he walked into the room**

1 ROUGE-1 Average_R: 0.07692
1 ROUGE-1 Average_P: 0.20000
1 ROUGE-1 Average_F: 0.11111

1 ROUGE-L Average_R: 0.07692
1 ROUGE-L Average_P: 0.20000
1 ROUGE-L Average_F: 0.11111
Experiments and Evaluation: Aspects

- Aspects:
  - Content
    - Automatic > a state-of-the-art method exist
  - Readability
    - ROUGE is not reliable for readability
    - Manual > There are not many automatic methods, mostly manual
Experiments and Evaluation: Readability Aspects by DUC

- Grammaticality
- Non-redundancy
- Referential clarity
- Focus
- Structure and coherence
Experiments and Evaluation: Survey

- Grammaticality
- Non-redundancy
- Referential clarity
- Focus
- Structure and coherence

First Survey
Experiments and Evaluation: Survey

- Grammaticality
- Non-redundancy
- Referential clarity
- Focus
- Structure and coherence

First Survey
Experiments and Evaluation: Survey

- Grammaticality
- Non-redundancy
- Referential clarity
- Focus
- Structure and coherence

First Survey

Second Survey
Experiments and Evaluation: Results

● Aspects:
  ○ **Content:**
    ■ Automatic > a state-of-the-art method exist
  ○ **Readability**
    ■ ROUGE is not reliable for readability
    ■ Manual > There are not many automatic methods, mostly manual

<table>
<thead>
<tr>
<th>Training Model</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>double-abstractive</td>
<td>2.15</td>
</tr>
<tr>
<td>ea-full-target</td>
<td>2.67</td>
</tr>
</tbody>
</table>
Experiments and Evaluation: Results

- Aspects:
  - Content:
    - Automatic > a state-of-the-art method exists
  - Readability:
    - ROUGE is not reliable for readability
    - Manual > There are not many automatic methods, mostly manual

<table>
<thead>
<tr>
<th>Training Model</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>double-abstractive</td>
<td>2.15</td>
</tr>
<tr>
<td>ea-full-target</td>
<td>2.67</td>
</tr>
<tr>
<td>ea-short-target</td>
<td>4.18</td>
</tr>
</tbody>
</table>
## Experiments and Evaluation: Comparison of Evaluation Aspects

<table>
<thead>
<tr>
<th>Training Model</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>double-abstractive</td>
<td>2.15</td>
</tr>
<tr>
<td>ea-full-target</td>
<td>2.67</td>
</tr>
<tr>
<td>ea-short-target</td>
<td>4.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ROUGE</th>
<th>double-abstractive</th>
<th>ea-full-target</th>
<th>ea-short-target</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1</td>
<td>0.23</td>
<td>0.29</td>
<td>0.54</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.16</td>
<td>0.21</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Observations : UNK Error

**Ground-truth Summary:** yesterday san francisco giants if barry bonds hit a 435-foot home run, his 756th, off a pitch from mike bacsick of the washington nationals, breaking the all-time career home run record, formerly held by hank aaron. the pitch, the seventh of the at-bat, was a 3-2 pitch, which bonds hit into the right-center field bleachers. matt murphy, a 22-year-old from queens in new york city, got the ball and was promptly protected and escorted away from the mayhem by a group of san francisco police officers.

**Computer Generated Summary:** yesterday san francisco giants [UNK] barry bonds hit a [UNK]home run, his 756th, off a pitch from [UNK] of the washington nationals, breaking the [UNK] home run in 1974.
Observations: Inaccurate Facts

Input
...
Hello Kitty was first introduced by Japanese company Sanrio in 1974. The cute round-faced ...

Ground-truth summary: the armband, which features "hello kitty" sitting on top of two hearts, will be worn by police officers who commit minor offences. These include, and parking in a prohibited area. The officers will also be forced to stay with the deputy chief all day in division office and will be forbidden to disclose their offences.

Computer-generated summary: the armband, which features "hello kitty" sitting on top of two hearts, will be worn by police officers who commit minor [UNK] include, and parking in a prohibited area. The officers will also be forced to stay with the deputy police in 1974.
Observations : Repetition

Computer-generated summary: A court in Kenya has sentenced a group of seven Somali pirates to five years each in jail, according to a statement by the European [UNK] mission EU [UNK] said that the men, “I have concrete proof that you attacked a vessel in the high seas and I order you to serve five years in prison seas and I order you to serve five years in prison seas and I order you to serve five years in prison seas and I order you to serve five years in prison seas and I order you to serve.”
Observations : Successful Summaries

Ground-truth summary: At least ten attackers with knives, dressed in black, attacked a train station in China yesterday. At least 28 victims were killed, with 113 more wounded by knives, Chinese state news agency reported. The local municipal government accuses "separatist forces" for the attack.

Computer-generated summary: At least ten attackers with knives, dressed in black, attacked a train station in China yesterday. At least 28 victims were killed, with 113 more wounded by knives, Chinese state news agency reported. The local municipal government accuses "separatist forces" for the attack.

Ground-truth summary: The United States Navy has successfully destroyed a crippled spy satellite in a decaying orbit, by intercepting it with a missile. A modified SM-3 missile was launched from the USS Lake Erie at 03:26 GMT this morning, and intercepted the USA-193 satellite around three minutes later. It has been reported that the satellite has broken into around 80 pieces, some of which have already re-entered the earth's atmosphere.

Computer-generated summary: The United States Navy has successfully destroyed a crippled spy satellite in a decaying orbit, by intercepting it with a missile. A modified SM-3 missile was launched from the USS Lake Erie at 03:26 GMT this morning, and intercepted the USA-193 satellite around three minutes later. It has been reported that the satellite has been damaged.
Recap, Conclusion & Future Work

- Contributions
  - One of the first large-scale multi-document summarization dataset for news domain
Recap, Conclusion & Future Work

- Contributions
  - One of the first large-scale multi-document summarization dataset for news domain
  - Wikisummarizer
Recap, Conclusion & Future Work

- Contributions
  - One of the first large-scale multi-document summarization dataset for news domain
  - Wikisummarizer
  - Unified multi-document summarization pipeline
Recap, Conclusion & Future Work

● Contributions
  ○ One of the first large-scale multi-document summarization dataset for news domain
  ○ Wikisummarizer
  ○ Unified multi-document summarization pipeline

● Conclusion
Recap, Conclusion & Future Work

● Contributions
  ○ One of the first large-scale multi-document summarization dataset for news domain
  ○ Wikisummarizer
  ○ Unified multi-document summarization pipeline

● Conclusion
  ○ Extractive Summarization proved to be a good method to transfer from multi-document to single-document
Recap, Conclusion & Future Work

● Contributions
  ○ One of the first large-scale multi-document summarization dataset for news domain
  ○ Wikisummarizer
  ○ Unified multi-document summarization pipeline

● Conclusion
  ○ Extractive Summarization proved to be a good method to transfer from multi-document to single-document
  ○ Better content selection resulted in better readability
Recap, Conclusion & Future Work

- Contributions
  - One of the first large-scale multi-document summarization dataset for news domain
  - Wikisummarizer
  - Unified multi-document summarization pipeline

- Conclusion
  - Extractive Summarization proved to be a good method to transfer from multi-document to single-document
  - Better content selection resulted in better readability
  - Even though there is a room for improvement, the idea behind the framework is promising
Recap, Conclusion & Future Work

● Contributions
  ○ One of the first large-scale multi-document summarization dataset for news domain
  ○ Wikisummarizer
  ○ Unified multi-document summarization pipeline

● Conclusion
  ○ Extractive Summarization proved to be a good method to transfer from multi-document to single-document
  ○ Better content selection resulted in better readability
  ○ Even though there is a room for improvement, the idea behind the framework is promising

● Future Work
Recap, Conclusion & Future Work

● Contributions
  ○ One of the first large-scale multi-document summarization dataset for news domain
  ○ Wikisummarizer
  ○ Unified multi-document summarization pipeline

● Conclusion
  ○ Extractive Summarization proved to be a good method to transfer from multi-document to single-document
  ○ Better content selection resulted in better readability
  ○ Even though there is a room for improvement, the idea behind the framework is promising

● Future Work
  ○ Extending the dataset
Recap, Conclusion & Future Work

● Contributions
  ○ One of the first large-scale multi-document summarization dataset for news domain
  ○ Wikisummarizer
  ○ Unified multi-document summarization pipeline

● Conclusion
  ○ Extractive Summarization proved to be a good method to transfer from multi-document to single-document
  ○ Better content selection resulted in better readability
  ○ Even though there is a room for improvement, the idea behind the framework is promising

● Future Work
  ○ Extending the dataset
    ■ Distant Supervision for clustering other datasets
Recap, Conclusion & Future Work

● Contributions
  ○ One of the first large-scale multi-document summarization dataset for news domain
  ○ Wikisummarizer
  ○ Unified multi-document summarization pipeline

● Conclusion
  ○ Extractive Summarization proved to be a good method to transfer from multi-document to single-document
  ○ Better content selection resulted in better readability
  ○ Even though there is a room for improvement, the idea behind the framework is promising

● Future Work
  ○ Extending the dataset
    ■ Distant Supervision for clustering other datasets
    ■ Merge with Multi-News
Recap, Conclusion & Future Work

● Contributions
  ○ One of the first large-scale multi-document summarization dataset for news domain
  ○ Wikisummarizer
  ○ Unified multi-document summarization pipeline

● Conclusion
  ○ Extractive Summarization proved to be a good method to transfer from multi-document to single-document
  ○ Better content selection resulted in better readability
  ○ Even though there is a room for improvement, the idea behind the framework is promising

● Future Work
  ○ Extending the dataset
    ■ Distant Supervision for clustering other datasets
    ■ Merge with Multi-News
  ○ Other setups of Wikisummarizer
Thanks for coming...
References


Image references


https://tr.pinterest.com/pin/451345193878807678/?lp=true


https://www.timeshighereducation.com/sites/default/files/styles/the_breaking_news_image_style/public/Pictures/web/n/c/o/numbers_on_podium.jpg?itok=-nVlhkPx