Background & Problem

- Document summarization is an important task with many real-world applications.
- Two primary modes of summarization: extractive and abstractive.

**Key Problems:**

- A lack of document-summary pairs.
- Few methods of fully unsupervised summarization exist, and of those that do, most are purely extractive.
- For multi-document summarization, datasets are few and far between.
- The length limit problem.
- ExtraPhrase: Succeeded at creating data that improved a transformer-based model's ability to generate summaries, but its dependency parsing-based system does not allow for control of output summary length.

Methods

- Data Augmentation Method: Two-step approach
  - TextRank
  - Back-Translation
- Baselines: heuristic for creating length-controlled abstractive pseudo-summary, and finetune BART on these summaries.

Solution

- ExtraPhraseRank: a data augmentation method that generates pseudo-summaries from documents with fine control over word count.
- Enables the fine-tuning of pre-training large language models without document-summary pairs.

Experiments

- We test fine-tuning a BART model on a variety of datasets
- XSum: 1k sampled news articles/summaries out of 200k
- Amazon: 1k review/synthetic summary pairs
- Evaluate ROUGE scores on XSum dataset across 7 experiments:
  - Given no gold summaries: no-fine-tuning, heuristic, TextRank, ExtraPhraseRank
  - Given some gold summaries: 1/3, 1/4, 1/5 synthetic, or only 1/5 gold
- Generate summaries with BART fine-tuned with ExtraPhraseRank for Amazon dataset and manually evaluate

Results and Conclusion

- On the XSum dataset, fine-tuning ExtraPhraseRank led to a modest improvement when no training data exists, outperforming heuristic and TextRank alone.
- Still performs poorly where gold summaries exist.
- On Amazon dataset, the summaries that BART generated after fine-tuning with synthetic data were not generally better.
- Conclusion: ExtraPhraseRank proved to be a mildly successful method of length-controlled data-augmentation in most cases, but it has significant limitations, and other methods should continue to be explored.

Table 1: Experimental Results on XSum Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-L</th>
<th>ROUGE-2</th>
<th>ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>27.39</td>
<td>14.021</td>
<td>34.091</td>
</tr>
<tr>
<td>No Fine-tuning</td>
<td>10.971</td>
<td>1.311</td>
<td>13.563</td>
</tr>
<tr>
<td>Baseline</td>
<td>10.539</td>
<td>1.046</td>
<td>14.655</td>
</tr>
<tr>
<td>TextRank</td>
<td>11.27</td>
<td>1.329</td>
<td>14.374</td>
</tr>
<tr>
<td>ExtraPhraseRank</td>
<td>12.008</td>
<td>1.482</td>
<td>14.972</td>
</tr>
<tr>
<td>ExtraPhraseRank &amp; 1/2 Gold</td>
<td>15.383</td>
<td>3.52</td>
<td>19.618</td>
</tr>
<tr>
<td>1/2 ExtraPhraseRank &amp; 1/2 Gold</td>
<td>15.994</td>
<td>24.202</td>
<td>30.708</td>
</tr>
</tbody>
</table>