**Overview**

**Motivation:**
- Automatic extraction of intelligence and useful information from data is one of the main goals in intelligence applications.
- The data extraction process still relies too much on manual effort which is slow, costly, and subject to human error.

**Problem:**
- Automatic text summarization system would enhance productivity and reduce errors
- However, most summarization systems require large amounts of high-quality labeled data
  - Lack of in-domain labeled data for target applications
  - Domain mismatch between state-of-the-art systems and target applications
- Unsupervised summarization approaches often produced varied results

**Approach:**
- Leverage large amount of unlabeled and in-domain data that are often under-utilized for building summarization systems
- Develop a weak learning pipeline to automatically create labels from unlabeled data
- Train a text summarization system using the generated labels

**Graph-Based Unsupervised Weak Learners**

- **Sentence Graph**
  - Consider each sentence to be a node, with an edge between each sentence pair
  - Edge weight is set to distance between sentence embeddings of incident nodes

- **Why graph-based methods?**
  - They’re unsupervised (a good choice for unlabeled dataset)
  - They take into account connections between all sentences to tell you which are important in the context of the given document
  - There are existing methods to find nodes that are important in the graph (high centrality or close to original graph)

**Dataset**
- **Train data:** KALIMAT: unlabeled 20,000 Arabic news articles [5]
- **Test data:** EASC: Labeled Arabic Summaries Corpus [8]
- 153 Arabic Wikipedia and news articles with 5 hand-written summaries a piece

**Results**

**Comparison to Unsupervised Approaches**

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid (XML-R Embedding)</td>
<td>17.87</td>
</tr>
<tr>
<td>First Sentence</td>
<td>13.68</td>
</tr>
<tr>
<td>Weakly Supervised BertSum</td>
<td>29.77</td>
</tr>
</tbody>
</table>

**Comparison to Supervised Approaches**

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centroid (XML-R Embedding)</td>
<td>24.43</td>
</tr>
<tr>
<td>Weakly Supervised BertSum</td>
<td>29.77</td>
</tr>
<tr>
<td>Qaroush et al. – Fully Supervised [11]</td>
<td>43.10</td>
</tr>
</tbody>
</table>

**Example Summary from Arabic Article**

- As for fuel demand, this sector produced most of the power needed in these countries from the beginning of the nineteenth century until the twentieth century.

**Conclusions/Future Work**

- A weak learning pipeline for summarization
  - Demonstrates improvement over unsupervised baseline system and cross-language supervised systems
  - Demonstrates promising performance vs. fully-supervised systems
- We plan to extend this pipeline to include additional weak learners
- Additional embeddings (TFIDF, average BERT) for centroid, TextRank
- PacSum graph centrality method [12]
- Graph compression [13]
- Perform experiments in a high-resource setting to directly compare supervised and weakly supervised summarization
  - Dataset: CNN/DailyMail [10] (or some high-resource language summarization corpus)
  - Split into traintest partitions

**TFIDF, Term-Frequency Inverse Document Frequency**