**ABSTRACTIVE SUMMARIZATION OF FACT CHECK REPORTS WITH PRE-TRAINED TRANSFORMER TUNING ON EXTRACTIVE SUMMARIES**

**AUTHOR:** Ing. Peter Vajdečka  
**SUPERVISOR:** prof. Ing. Vojtěch Svátek, Dr.

---

**MOTIVATION AND GOALS**

Fact-checking reports are frequently too long for a casual reader, and contain auxiliary parts not directly relevant for judging the claim veracity. Automated creation of fact check report summaries is thus a topical task.

**Goals:**
- Review modern approaches to automated text summarization and identify methods suitable for generating a summary of a fact check report as a particular kind of document.
- Gather data from fact-checking sites, in particular, demagog.cz for Czech and politifact.com for English (scraped).
- Propose hybrid summarization model to be comparable with state-of-the-art summarizing models in the fact-checking domain.

---

**RELATED WORKS**

Our proposed method was compared to the following works in terms of the ROUGE metric (or ROUGE RAW – a language independent metric):

- On the Politifact dataset:
  - Atanasova et al. [1] explored a supervised BERT-based technique for jointly predicting the truth of a claim and extracting supporting explanations from fact-checked claims.
  - Kazemi et al. [2] utilized GPT-2 for abstractive summarization and Biased TextRank for extractive summarization as alternative approaches.

- On the SumeCzech dataset (not fact-checking focused, but the only summarization dataset available for Czech):
  - Straka et al. [3] applied different extractive techniques and proposed the ROUGE RAW metric for the Czech language.
  - Marek et al. [4] came up with named entities summarization, even improving results of [3].

---

**PROPOSED HYBRID METHOD**

The core contribution of our hybrid approach is extroactive summarization based on Local outlier factor (LOF) and sentence representation by fine-tuned Sentence-BERT. The input pair consists in a fact check report and a justification (as its manually created summary). Sentence-splitting is applied to the report, and embedding of separated sentences are created using alternative methods such as TF-IDF, DOC2VEC or Sentence-BERT; it was fine-tuned Sentence-BERT that returned the best vector representation. We then compute the normalized LOF. Sentences above a certain LOF threshold (optimized during the experiments) are removed, and the remaining ones become an extractive summary of the report, which together with its corresponding justification enters the T5 transformer to generate abstractive summaries.

---

**EVALUATION RESULTS**

Table 1: Politifact results

<table>
<thead>
<tr>
<th>System</th>
<th>T5 Baseline</th>
<th>Claim+T5 Baseline</th>
<th>Claim+BERT (LOF+TF-IDF+LOF+T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE 1</td>
<td>11.3</td>
<td>11.6</td>
<td>18.1</td>
</tr>
<tr>
<td>ROUGE 2</td>
<td>6.9</td>
<td>7.1</td>
<td>13.1</td>
</tr>
<tr>
<td>ROUGE L</td>
<td>11.8</td>
<td>12.1</td>
<td>18.7</td>
</tr>
</tbody>
</table>

Table 2: Demagog results (all for variants of present work)

<table>
<thead>
<tr>
<th>System</th>
<th>Score</th>
<th>Feature 1</th>
<th>Feature 2</th>
<th>Feature 3</th>
<th>Feature 4</th>
<th>Feature 5</th>
<th>Feature 6</th>
<th>Feature 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumeCzech</td>
<td>Score</td>
<td>Feature 1</td>
<td>Feature 2</td>
<td>Feature 3</td>
<td>Feature 4</td>
<td>Feature 5</td>
<td>Feature 6</td>
<td>Feature 7</td>
</tr>
<tr>
<td>T5 Baseline</td>
<td>35.7</td>
<td>17.3</td>
<td>32.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5 Baseline</td>
<td>35.1</td>
<td>17.2</td>
<td>32.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5 Baseline</td>
<td>35.4</td>
<td>17.3</td>
<td>32.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: SumeCzech results

---

**CONCLUSION**

- On the ROUGE metrics, results for the proposed hybrid summarization approach outperform previous studies [1, 2, 3, 4] on all three data sets (see Tables 1, 2 and 3).
- We plan to extend the experiments to other domains and datasets (different from fact-checking) and aim to improve the ranking of input sentences by applying other features beyond the LOF score.

---

**REFERENCES**


