Hybrid Long Document Summarization using C2F-FAR and ChatGPT: A Practical Study

Guang Lu, Sylvia B. Larcher, Tu Tran
July 11, 2023

SwissText Conference 2023
AGENDA

1. Introduction
2. Related Work
3. Data and Methods
4. Results and Discussion
5. Conclusions and Outlook
6. QA
1. Introduction

Text Data from Long Documents

Data Preprocessing and Cleaning

Business Articles (20-30 Pages)

Business Books (> 100 Pages)

Hybrid Summarization

Impact on Long Document Summarization in Business
1. Introduction

The contributions of this work:

- We presented the first study using a hybrid text summarization pipeline with C2F-FAR and ChatGPT for summarizing long documents.

- We studied the performance of ChatGPT in summarizing long business articles and books highlighted by automated evaluation metrics and human annotations.

- We provided insights into the strengths and weaknesses of the proposed summarization pipeline and pointed out the next key challenges that LLMs like ChatGPT need to solve in summarizing long documents.

Many summarization models have been developed to summarize long documents of up to 10,000 words:

- Trained based on given datasets and need to be fine-tuned or completely re-trained for new text domains
- The application of these models in practice is challenging also due to the complex nature of the models
- The zero-shot learning capability of LLMs is seen as a promising alternative for tasks of summarizing long documents
3. Data and Methods

Based in Switzerland, getAbstract AG is a world-renowned company whose primary mission is:

- Provide high-quality, expertly written summaries for a wide range of English-language business books, articles, videos and business reports
- Help readers quickly acquire new knowledge in the business world

We used business articles and books and a subset of getAbstract summaries written by humans for summary quality evaluation.

We randomly selected and cleaned 20 articles and 20 books to test the proposed hybrid summarization pipeline.

We implemented two methods to prepare the cleaned text for the summarization.

- The first method is based on fully automated tools such as Python and Parse.¹
- The second method is based on manual and automatic cleaning with Python.

The 20 articles have an average word count of 5,603 and the 20 books have an average word count of 67,031.

¹ https://github.com/axa-group/Parsr
C2F-FAR (Coarse-to-Fine Facet-Aware Ranking) is an unsupervised method that can automatically select sentences from a document by considering the semantic similarity between different blocks of text and between different sentences. The advantage is that this method can be applied directly without having to be trained on a large amount of data.

However, there are several parameters that can be tuned for optimal performance on the targeted dataset. The parameters can affect, for example, the number of semantic blocks, the threshold above which the sentences in a block must be selected, and the output length of the summary.
3. Data and Methods

We used ChatGPT for an abstractive summarization based on the extractive summaries from C2F-FAR. We designed the prompts as a simple two-step process:

- For business articles, we asked ChatGPT to “Please summarize the following text in your own words in about 25 sentences.” If the summaries are too short, we further asked ChatGPT to generate 25 sentences or add more details to the summary.

- For business books, we repeated the above procedure, but had to enter the sentences extracted from C2F-FAR into ChatGPT in chunks (about three pages in a Word document) so that ChatGPT could summarize step by step.

"Ten Economic Facts about Immigration" by Michael Greenstone and Adam Looney presents a comprehensive examination of the economic impact of immigration on the United States. The authors draw on a range of data and research to provide a well-rounded perspective on the topic and present their findings in the form of ten key facts. The central argument of the book is that immigration is an economic benefit for the United States, and the authors provide a wealth of evidence to support this claim.

The authors begin by providing historical context for immigration in the United States, highlighting its role in the country’s economic growth. They then proceed to present the ten key facts about immigration. The first of these is that immigrants are more likely to start businesses than native-born Americans. This is an important fact as it highlights the entrepreneurial spirit of immigrants and their potential to contribute to economic growth.

The second key fact is that immigrants are responsible for a significant portion of...
C2F-FAR can be used as an extractive basis for the hybrid summarization pipeline.

### Table 1: Comparison of C2F-FAR generated and human-written summaries with reference-based metrics. “Avg” stands for the average and “Std” stands for the standard deviation. “B” stands for the business articles and “E” stands for the economic reports.

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-L</th>
<th>BERTScore</th>
<th>Words (Human)</th>
<th>Words (Model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg (B)</td>
<td>0.45</td>
<td>0.10</td>
<td>0.17</td>
<td>0.84</td>
<td>363</td>
<td>560</td>
</tr>
<tr>
<td>Std (B)</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>28</td>
<td>67</td>
</tr>
<tr>
<td>Avg (E)</td>
<td>0.44</td>
<td>0.11</td>
<td>0.16</td>
<td>0.84</td>
<td>605</td>
<td>638</td>
</tr>
<tr>
<td>Std (E)</td>
<td>0.06</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>503</td>
<td>178</td>
</tr>
</tbody>
</table>

### Table 2: Comparison of C2F-FAR generated and human-written summaries with reference-free metrics. The definitions of “Avg”, “Std”, “B”, and “E” are the same as in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>BLANC help (Human)</th>
<th>BLANC help (Model)</th>
<th>BLANC tune (Human)</th>
<th>BLANC tune (Model)</th>
<th>ESTIME alarms (Human)</th>
<th>ESTIME alarms (Model)</th>
<th>ESTIME soft (Human)</th>
<th>ESTIME soft (Model)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg (B)</td>
<td>0.11</td>
<td>0.19</td>
<td>0.06</td>
<td>0.15</td>
<td>187</td>
<td>26</td>
<td>0.61</td>
<td>0.97</td>
</tr>
<tr>
<td>Std (B)</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td>26</td>
<td>6</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>Avg (E)</td>
<td>0.12</td>
<td>0.17</td>
<td>0.05</td>
<td>0.14</td>
<td>292</td>
<td>24</td>
<td>0.65</td>
<td>0.98</td>
</tr>
<tr>
<td>Std (E)</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.05</td>
<td>262</td>
<td>9</td>
<td>0.04</td>
<td>0.01</td>
</tr>
</tbody>
</table>
4. Results and Discussion

C2F-FAR Summary

Millennials are the first generation to grow up immersed in a digital world. While there are some distinctions among the generations, Millennials attitudes are not poles apart from other employees. We’ve also uncovered three uncomfortable truths that apply to employees of all ages. Lastly, we’ve made five practical recommendations for helping a multigenerational workforce thrive in today’s volatile work environment. Millennials have similar career aspirations to those of other generations. Surprisingly, it’s largely Gen X employees, not Millennials, who think everyone on a successful team should be rewarded. Myth 3: Millennials are digital addicts who want to do and share everything online, without regard for personal or professional boundaries. Millennials are also quite capable of distinguishing between the personal and professional realms and exercising discretion when they use social media. Many organizations encourage their employees to leverage their personal networks, but this approach could increase the potential for misuse or mistakes if employees don’t have the proper guidance. Myth 4: Millennials, unlike their older colleagues, can’t make a decision without first inviting everyone to weigh in. Despite their reputation for crowdsourcing, Millennials are no more likely than their older colleagues to solicitation advice at work. True, more than half of all Millennials say they make better business decisions when a variety of people provide input. Baby Boomers, by contrast, feel far less compelled to include others or worry about seeking consensus and are more skeptical about whether the boss knows best (see Figure 7). Baby Boomers accustomed to making decisions on their own may find it difficult to shift to a more collaborative culture, which can cause tension between older and younger employees. Having the aptitude and tools to do this quickly is essential, as the business landscape becomes more interconnected and complex.

Another fiction. When Millennials change jobs, they do so for much the same reasons as Gen X and Baby Boomers. But, as Figure 8 shows, there are no overwhelming generational differences. Seventy-five percent of Millennial respondents said they’ve held their current positions for three years or more, suggesting that they are no more inclined than older colleagues to gallivant from one job to the next. In the course of our research, we identified three insights that apply universally and should give business leaders everywhere cause for concern. More than half of the people we surveyed don’t fully understand key elements of their organizations strategy, what they’re supposed to do or what their customers want. While up to 60 percent of Gen X respondents believe they have a good grasp of these fundamentals, many of their colleagues are struggling (see Figure 9). We asked our respondents to rate their organizations leaders on a number of criteria. We analyzed their answers by role, as well as by generation, to find out how job status influences their perceptions (see How we defined each role). Gen X leaders, in particular, overrate how well they inspire confidence and recognize employees accomplishments (see Figure 10).

Human Summary

Millennials those aged 21 to 34 in 2015 are the first generation to grow up fully immersed in technology. These digital natives are expected to revolutionize the workforce; older people see them as different from previous generations. Much fear and misconception surrounds millennials; one of the most unmistakable stereotypes is that they are lazy, entitled, selfish and shallow. However, research tells a different story. Consider five millennial myths, debunked:

- Their career goals are distinct from those of older generations. Like gen X and baby boomers, however, millennials desire financial security and seniority and believe in the value of inspirational leadership and performance-linked recognition. They expect endless praise for their accomplishments. In truth, millennials value ethics and fairness more. They are digital addicts who do everything online. Although millennials are good at interacting online, they prefer learning new work-related skills face-to-face. They make decisions by inviting others to pitch in. While 50% of millennials report valuing group input, they are no more likely than any other generational group to cross-reference advice. They jump jobs if their current work doesn’t fulfill their passions. In reality, millennials typically change jobs for the same reason as other generations. Millennials as digital natives bring vital value to a work environment in the midst of a digital revolution. But in many ways, they are a lot like their older colleagues. Firms must recognize some worrying truths about all employees, regardless of generation. Many employees don’t understand their company’s strategy, what they’re supposed to do, or what their customers want. Leaders need to find creative ways to connect with employees. Most employees believe that their companies handle the customer experience ineffectively and embrace new technologies slowly. Companies should assess the risks of not adopting changes and should introduce innovations to enhance the customer experience. As more millennials have embarked on their careers, expectations of a technological revolution in the workplace have increased. Business leaders must recognize that employees of all ages are complex individuals working in an environment that’s becoming more virtual, more diverse and more volatile by the day and initiate suitable changes in the workplace.
We also had ChatGPT directly summarize some English articles and books in the dataset and examine the quality of the summaries using human evaluation. The prompt reads, “Please summarize the article (name of article) in approximately 500 words using a (description of style) style.”

- Long summaries generated very quickly
- Very mechanical and sometimes repetitive structure
- Style controlled to some extent
- Content sometimes somewhat general
- Faithfulness problems
- Wrong document summarized
4. Results and Discussion

We tested ChatGPT’s paraphrasing ability on the input texts.

For business articles, the prompt was “Please rewrite the text in your own words”. Summaries written by humans were considered the golden standard.

For business books, the prompt was “Please summarize (rewrite, summarize and rewrite) the text in your own words in about 25 sentences”. Summaries formed from the top 25 sentences using C2F-FAR were considered the gold standard.
To complete the evaluation of the summaries produced by our hybrid summarization pipeline, we conducted thorough human evaluations and even compared the summaries produced by GPT-3.5 and GPT-4.

- The summaries written by humans are much longer than those produced by ChatGPT. At least a two-step query is required to ensure that ChatGPT reproduces the content in sufficient detail. Because the input to ChatGPT is long, smaller chunks are required to divide the long document. **However, due to the very mechanical structure of the text generated by ChatGPT, the coherence between the texts generated by different chunks is not ideal.**

- The summaries produced by ChatGPT appear to be correct at the beginning. **Halfway through the text, the machine starts to hallucinate more and more.** Other problems include a serious lack of coherence within the paragraph, illogical conclusions, redundancies, literal repetitions and inexplicable abbreviations.

- As far as style is concerned, **the texts created with ChatGPT do not contain subheadings, quotations, metaphors, idiomatic expressions or puns that make the text unique and interesting.** Only the text created with GPT-4 has some of these stylistic features.
Conclusions:

- The results measured against the existing automated evaluation metrics suggest that the machine-generated summaries look as good as those written by humans, especially when the ROUGE-1 score is taken into account.

- Several critical issues were identified in the human evaluation in terms of text coherence, faithfulness and style. Most importantly, models as powerful as ChatGPT and GPT-4 can still produce unfaithful summaries given the input texts that need to be carefully checked by human editors before they are used in practice.

- This shows that existing LLMs cannot yet fully replace the role of human effort in composing accurate and correct texts, especially considering that there is no effective way to detect hallucinations in texts.

Outlook:

- The rapid development of the GPT family models provides faster, better and more inspiring results in text summarization compared to the other state-of-the-art models.

- One important work is to focus on how to effectively and efficiently assess the quality of long document summaries in the production environment and how to use the model-generated texts to improve human work efficiency.

- When using ChatGPT services, privacy issues in real applications must also be considered.
Thank you for your attention! QA?