



Figure 2: Backend pipeline of SUMMPILOT.

Method: SUMMPILOT

As shown in Figure 2, SUMMPILOT employs multiple internal modules to generate and customize summaries. By leveraging LLMs (specifically GPT-4o (OpenAI 2024)), our system achieves significant flexibility through well-designed prompts and advanced language comprehension.

Automatic Summarization When multiple documents are provided as input, the automatic summarization module utilizes an LLM to generate a coherent abstractive summary.

Relation Extraction and Entity Clustering The relation extraction module identifies relational triples from input documents, forming the basis of a semantic graph that visually represents the information. SUMMPILOT’s graph visualization maps these triples with directed edges to indicate relationships and adjusts node sizes to emphasize key entities. This helps users better understand the content structure and precisely control their information needs. After generating the graph, entity clustering groups mentions of the same entities. For example, the triples $\langle \text{Tom-is married to-Jane} \rangle$ and $\langle \text{Tom’s wife-aged-30} \rangle$ are combined into $\langle [\text{Tom’s wife, Jane}] \text{-aged-30} \rangle$. This clustering clarifies variations in expressions referring to the same concept. A representative entity, determined by frequency, is then assigned to each subject and object for visualization and interaction.

Interactive Summarization The interactive summarization module generates customized summaries based on user needs. This module interprets user requests and incorporates them into prompts for the LLM, addressing inclusion/exclusion criteria and user-defined preferences in both structured and flexible contexts. This adaptability supports iterative refinement, enhancing user engagement and satisfaction.

Summary Evaluation The evaluation of generated summaries involves three components: compression, coverage, and factual consistency. Compression measures how much the summary condenses the original text, while coverage assess how well the summary retains information from the source documents (Grusky, Naaman, and Artzi 2018). Factual consistency, adapted from FactScore (Min et al. 2023) and customized for our task, evaluates the accuracy by decomposing each summary into atomic facts and verifying

Aspect	Score
Semantic Graph Effectiveness	4.4 (0.8)
Entity Clustering Effectiveness	3.8 (0.7)
Intuitiveness Relation Extraction	5.0 (0.0)
Ease of Control	4.6 (0.5)
Possible Error	4.6 (0.5)
Effectiveness of Including Requirements	4.8 (0.4)
Effectiveness of Excluding Requirements	4.8 (0.4)
Summary Satisfaction	4.8 (0.4)
Summary Trustworthiness	4.4 (0.8)

Table 1: Result of usefulness questionnaire. The scores are reported as the average (standard deviation).

them against the original documents. The final factual consistency score is the proportion of verified atomic facts. Unverified facts are flagged as “possible errors” in the interface.

User Studies & Evaluations

The results in Table 1 highlight several key advantages of SUMMPILOT. Participants appreciated the semantic graph for enhancing document understanding and praised the intuitive interaction enabled by relation extraction. The system’s error detection feature was also recognized as highly beneficial for evaluating generated summaries. These findings underscore SUMMPILOT’s strengths in facilitating comprehensive document analysis, enabling precise user interaction, and supporting effective summarization and evaluation.

Conclusions & Future Work

SUMMPILOT is an interaction-driven, customizable summarization system designed for flexibility and effectiveness, as demonstrated by user studies. It establishes a robust foundation for efficiently summarizing complex business documents and news articles, making it particularly beneficial for reporters and students who need to extract essential information. Our demo video is available at (<https://youtu.be/7ZUBIyqPpbs>). Future work will focus on multilingual support and the integration of multimodal LLMs to handle various data types, such as images and videos, enhancing user control and improving human-computer interaction.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1C1C1008534), and Institute for Information & communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 20210-01341 (Artificial Intelligence Graduate School Program, Chung-Ang University).

References

- Adenuga, I. J.; Hanrahan, B. V.; Wu, C.; and Mitra, P. 2022. Living Documents: Designing for User Agency over Automated Text Summarization. In *Extended Abstracts of CHI*, 1–6.
- Grusky, M.; Naaman, M.; and Artzi, Y. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of NAACL*, 708–719.
- Hirsch, E.; Eirew, A.; Shapira, O.; Caciularu, A.; Cattan, A.; Ernst, O.; Pasunuru, R.; Ronen, H.; Bansal, M.; and Dagan, I. 2021. iFacetSum: Coreference-based Interactive Faceted Summarization for Multi-Document Exploration. In *Proceedings of EMNLP (Demo Track)*, 283–297.
- Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of EMNLP*, 12076–12100.
- OpenAI. 2024. Hello GPT-4o. Accessed: May 21, 2024.
- Slobodkin, A.; Nachum, N.; Amar, S.; Shapira, O.; and Dagan, I. 2023. SummHelper: Collaborative Human-Computer Summarization. In *Proceedings of EMNLP (Demo Track)*, 554–565.
- Zhang, X.; Li, J.; Chi, P.-W.; Chandrasegaran, S.; and Ma, K.-L. 2023. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In *Proceedings of CHI*, 1–16.