Utilising Transformer Models for Controllable Scientific Abstractive Summarization

Sebastian Weidinger, Sarah Frank,
Andreas Wagner, Christian Gütl

Open Search Symposium 2024 - October 10, 2024

Introduction

- Growing number of scientific publications
  → **information retrieval** challenges
- Need for **efficient summarization** tools
  - Complex terminology
  - Long text
- **Key challenges**:
  - Domain specificity
  - Computational cost
  - Traceability

Problem Statement

- **Traditional short-text models** are insufficient for scientific texts
  - E.g. Only for news
- Scientific articles **average 10.7k tokens** vs. 1k tokens for news
- High **computational costs**
- Large **input size**
- **Accuracy and traceability** are crucial for scientific summaries

Motivation

- **Efficient model**
  - Low computational costs
  - Affordable performance and quality
- **Length controllability**
- **Sufficient context size**
  - > 11k tokens input size
- **Capture scientific wording**
  - Close to human-written text

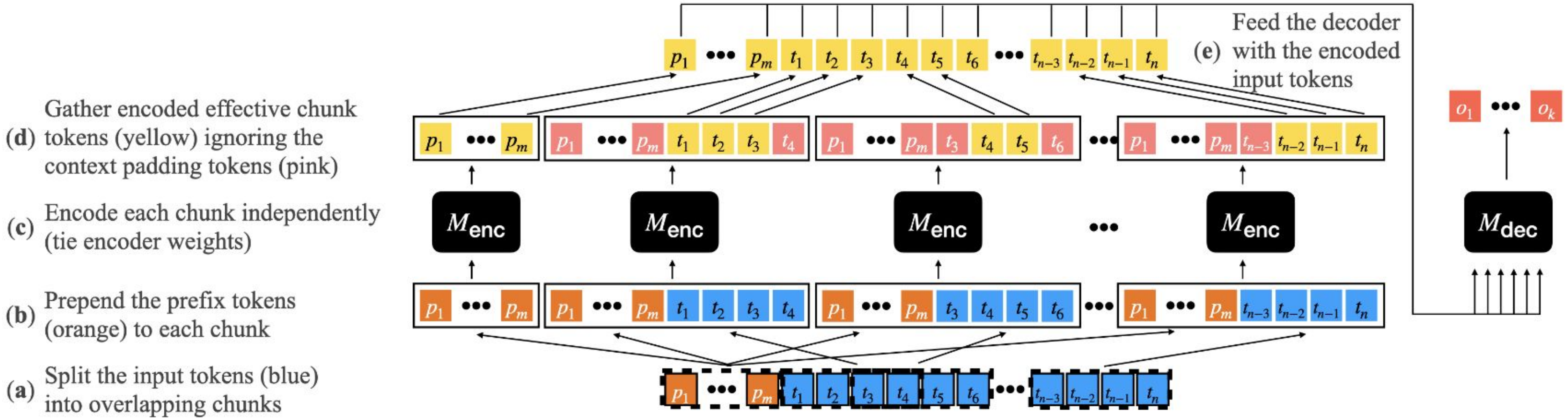
Dataset Creation

- Created a new dataset:
  - **OpenReview Contribution** (1.7k)
  - Scraping OpenReview.net
    - Open-access platform for peer review
- Focus on **computer science** papers
  - e.g. NeurIPS, ICLR
- Multiple summary lengths of different reviewers
  - **Controllable summarization**
    - 7 summary lengths
  - **Human-written summaries** as gold standard

Proposed Solution - Model

- **SLED** (Ivgi, Shaham & Berant, 2022)
  - Short-Length Encoder Decoder
  - **Efficient processing**
    - **Fusion-in-Decoder**
    - Divides text into **manageable chunks**
    - **Processes chunks** separately
    - Merged in decoding step
  - Balances **performance** and **computational cost**
  - **Increase input size** e.g. 16k tokens
- Leverages short-text pretrained LMs
  - Context size of 1k tokens and ~139M parameters

SLED - Architecture

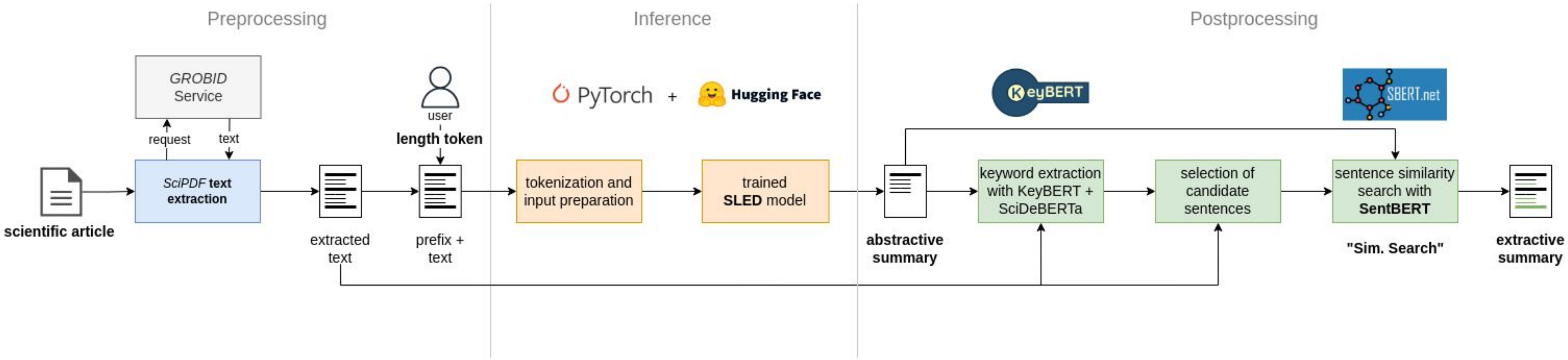


ISDS

Proposed Solution - Architecture

- **Preprocessing**
  - Text extraction from scientific papers
  - Add length prefix
- **Inference**
  - SLED model
  - Generation of abstract summary
- **Postprocessing**
  - Similarity Search (“Sim. Search”)
    - SentBERT (Reimers & Gurevych, 2019)
  - Generation of extractive summary

System Architecture



Experimental Setup - Model

- **SLED**
  - Comparison to extractive and abstractive methods
    - E.g. TextRank, BART and GPT-3.5
    - **Performance** and **quality aspects**
- SLED advantage
  - Fusion-in-decoder
    - **Long-range dependencies**
    - **Low computational costs**
  - **Input size of 12k** tokens with hardware settings used
    - RTX 3070
    - Memory 8 GB GDDR6

Experimental Setup - Model

- **“Sim. Search”**
  - Similarity search based on semantic search
  - Introduce simple traceability
  - Comparison to other extractive methods

Experimental Setup - Metric

- **Performance** comparison
  - **ROUGE** (Lin, 2004)
    - Lexical-based metric
  - **BERTScore** (Zhang, Luan, & Liu, 2019)
    - Semantic-based metric
- **Quality** comparison
  - **UniEval** (Liu & Liu, 2021)
    - Multidimensional deep learning-based evaluator
    - Automatic evaluation
    - **Coherence, factual consistency, fluency and relevance**

Experimental Setup

- Baseline models for **performance** comparison
  - **BART** (Lewis et al., 2020)
    - 1k max. input tokens
  - **TextRank** (Mihalcea & Tarau, 2004)
    - Extractive method
- Additional **GPT-3.5-turbo** for **quality** comparison
  - Max. input size of 16k
- Dataset
  - OpenReview Contribution
    - Full or subset

Results & Findings - Performance

- SLED performance
  - Outperforms baseline models on **long scientific documents**
- Better results with **controllable summary lengths**
- High **similarity** to human-crafted summaries (BERTScore)

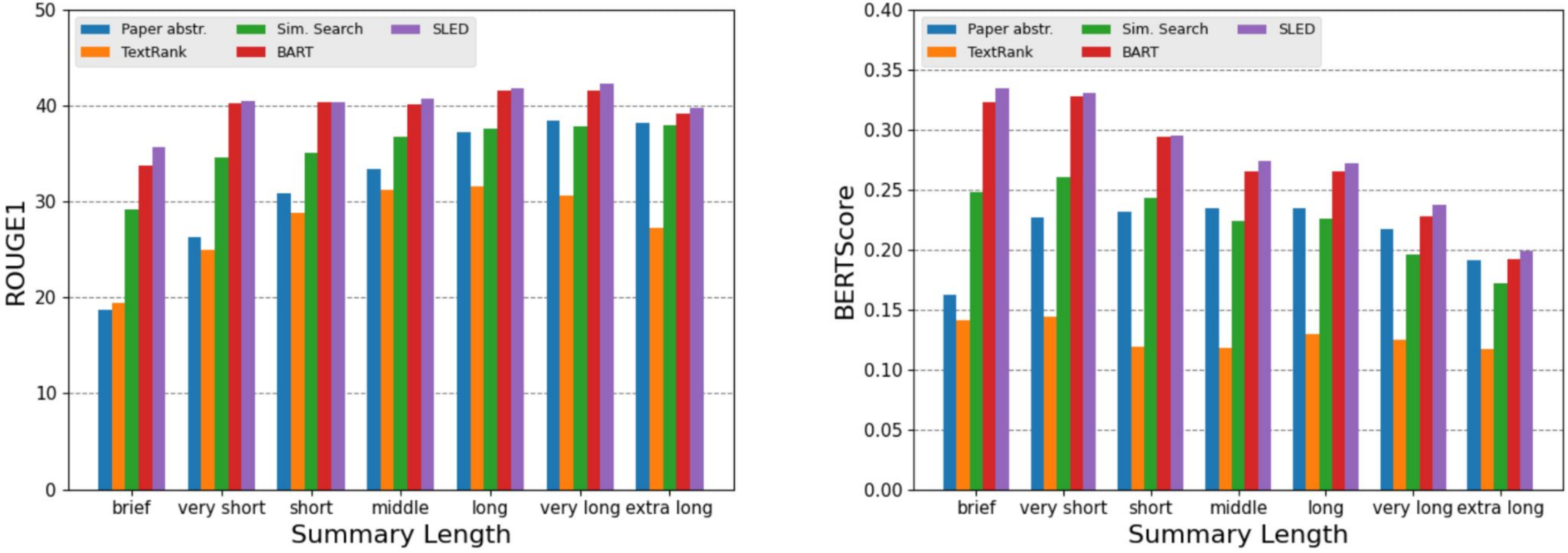
Performance

Comparison to human-written summaries

Method	**Input source**	**Length signal**	**ROUGE1**	**ROUGE2**	**ROUGELsum**	**BERTScore**
heuristic	paper abstr.	no	32.73	9.39	20.23	0.220
TextRank	full paper	no	29.09	6.21	19.26	0.114
TextRank	full paper	yes	30.95	6.52	20.41	0.128
Sim. Search	summ.+paper	yes	35.77	9.60	23.44	0.229
$\text{BART}_{\text{base}}$	1K tokens	yes	36.81	10.45	33.06	0.276
$\text{SLED}_{\text{base}}$	12K tokens	no	32.68	9.90	29.40	0.268
$\text{SLED}_{\text{base}}$	12K tokens	yes	**36.95**	**10.81**	**33.12**	**0.282**

ISDS

Performance



(a) Comparison on ROUGE1.

(b) Comparison on BERTScore.

Results & Findings - Quality

- SLED **comparable results** to GTP-3.5 on quality
- **SLED** higher performance in **fluency** and **relevance** compared to human written texts
- Affordable performance

Quality

Comparison to human-written summaries

Method	Type	#Params	Coherence	Consistency	Fluency	Relevance	Average
paper abstr.	extr.	-	**94.19**	**94.35**	88.80	85.42	**90.69**
TextRank		-	40.36	68.28	76.71	35.82	55.29
Sim. Search		-	61.55	82.91	87.55	55.21	71.80
$GPT_{zero\text{-}shot}$	abstr.	~20B	92.37	84.47	**91.63**	**91.52**	90.00
$BART_{base}$		139M	90.22	82.84	86.11	86.81	86.49
$SLED_{base}$		139M	89.08	80.99	88.93	87.54	86.64

ISDS

Key Insights

- **Model efficiency**
  - SLED offers an efficient approach with **affordable performance** vs. larger models (GPT-3.5)
- **Length control**
  - Enhances **summarization accuracy**
- **Semantic Search**
  - Improves reliability by identifying **original sentences**
  - Provides a good extractive summary

Conclusion & Future Work

- **Conclusion**
  - Dataset demonstrates high quality
  - SLED is a strong option for long-document summarization balancing performance and cost
- **Future Work**
  - Improve factual consistency and explore other efficient approaches

Q&A

Questions?

References

Ivgi, M., Shaham, U., & Berant, J. (2022). *Efficient Long-Text Understanding with Short-Text Models*.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics.

Zhang, T., Luan, Y., & Liu, J. (2019). BERTScore: Evaluating text generation with BERT. *In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 612-621). ACM.

Liu, Y., & Liu, S. (2021). UniEval: A Unified Framework for Text Generation Evaluation. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 5310–5320). Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Marjanovic, M., Stiennon, N., & Lowe, R. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 404–411). Association for Computational Linguistics.