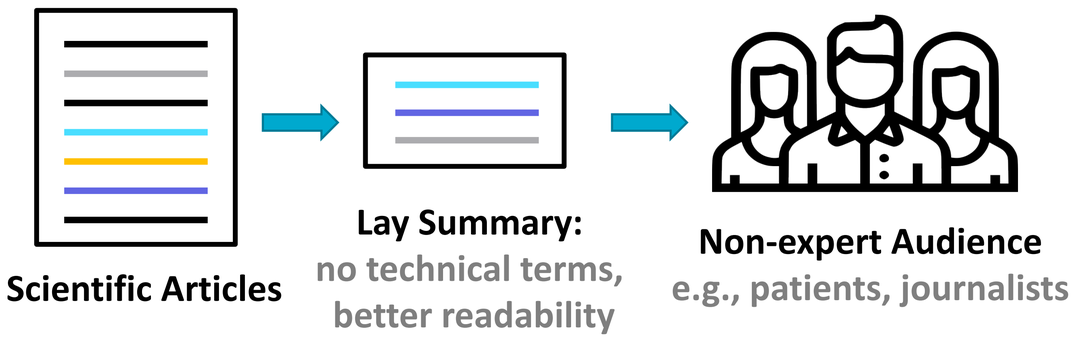
CSIRO DATA61 TEAM AT BIOLAYSUMM: LAY SUMMARISATION OF BIOMEDICAL RESEARCH ARTICLES USING GENERATIVE MODELS

Mong Yuan Sim, Xiang Dai, Maciej Rybinski, Sarvnaz Karimi
CSIRO Data61, The University of Adelaide

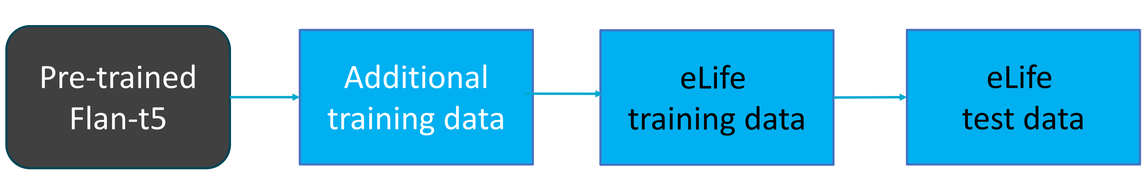


BioLaySumm 2023 Shared Task

- Abstractive summarization of biomedical articles, with an emphasis on controllability and catering to non-expert audiences.
- **Datasets:** From two biomedical journals PLOS and eLife
- **Input:** An article's abstract and main text
- **Output:** Generate a lay summary.
- **Metrics:**
  - **Relevance**: ROUGE (1, 2, and L) and BERTScore
  - **Readability**: Flesch-Kincaid Grade Level (FKGL) and Dale-Chall Readability Score (DCRS)
  - **Factuality**: BARTScore
- CSIRO Data61 Team **ranked 2nd for the relevance** criteria and **3rd overall** among 21 competing teams.

Take Away #1: GPT-3.5 can be used to generate additional training data to improve model effectiveness

- We investigate using **intermediate task pre-training** to improve model effectiveness. The FLAN-T5 model is first fine-tuned on additional training data, then on eLife training data, and finally evaluated on eLife test data. Note that eLife training set is small.



- We used the following additional training data:

a. The Extreme Summarization (XSum) dataset: consisting of news articles accompanied with one-sentence summary

b. PLOS: scientific journal publication, very similar to eLife

c. GPT-3.5 (P) – paraphrase eLife's gold summary using GPT-3.5. Prompt: *"Paraphrase the following paragraph: [Summary]*

d. GPT-3.5 (S) – summarise eLife's source document using GPT-3.5. Prompt: *"Summarise the following document using plain text: [Input Document]*

	#	ROUGE-1	ROUGE-2	ROUGE-L	BERTS.	FKGL	DCRS	BARTS.
				eLife Dev				
eLife	4346	0.495	0.146	0.469	0.855	**9.783**	**8.170**	-2.406
PLOS	24773	0.342	0.075	0.314	0.835	15.364	11.553	-2.421
Combined	29119	0.449	0.126	0.426	0.846	10.334	8.175	**-2.223**
PLOS → eLife	24773 + 4346	0.503	**0.152**	0.478	0.856	9.918	8.237	-2.415
XSum → eLife	203017 + 4346	0.495	0.147	0.470	0.855	9.866	8.176	-2.398
GPT-3.5 (P) → eLife	4346 + 4346	**0.510**	**0.152**	**0.484**	**0.857**	9.904	8.196	-2.528
GPT-3.5 (S) → eLife	4346 + 4346	0.502	0.151	0.477	0.856	10.052	8.215	-2.439
				eLife Test				
GPT-3.5 (P) → eLife‡	4346 + 4346	0.489	0.130	0.463	0.855	10.013	8.316	-2.612

Table 2: The impact of training data on the effectiveness. A -> B indicates sequential transfer learning where FLAN-T5-xl model is fine-tuned on A training data and then on B. # shows the number of training examples. ‡ indicates our final submission.

- Additional training data helps to improve relevance scores (ROUGE and BERTScore). GPT-3.5 (Paraphrasing summary) outperforms others despite its smaller size than PLOS and XSum.

Take Away #2: Larger models can generate summaries with higher relevance but not necessarily of higher readability

- We compare three FLAN-T5 versions: base 250M, large 780M, xl 3B
- Larger models produce lay summary with high relevance scores but lower readability and factuality scores.
- FLAN-T5-base achieves the best factuality and readability score (DCRS) when evaluating on eLife and PLOS development sets.

Dataset	Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	FKGL	DCRS	BARTScore
eLife Dev	FLAN-T5-base	0.411	0.108	0.393	0.834	10.148	**7.100**	**-2.101**
	FLAN-T5-large	0.469	0.136	0.446	0.849	**9.700**	7.850	-2.213
	FLAN-T5-xl	**0.495**	**0.146**	**0.469**	**0.855**	9.783	8.170	-2.406
PLOS Dev	FLAN-T5-base	0.493	0.186	0.455	0.863	15.132	**11.088**	**-1.882**
	FLAN-T5-large	0.497	0.187	0.459	0.864	14.948	11.163	-1.894
	FLAN-T5-xl	**0.502**	**0.190**	**0.462**	**0.865**	**14.826**	11.194	-1.908
eLife Test	FLAN-T5-xl	0.480	0.130	0.454	0.854	9.804	8.224	-2.493
PLOS Test ‡	FLAN-T5-xl	0.497	0.194	0.460	0.867	15.089	11.372	-1.887

Table 1: The impact of model size on effectiveness. Base model has 250M parameters; large has 780M parameters; and xl has 3B parameters. ‡ indicates our final submission.

Take Away #3: Truncating long input document has a high impact on the relevance but not on the readability and factuality

- Our model generates a highly relevant summary when the input document length is between 512 (2^9) and 1024 (2^{10})
- Relevance scores decrease when input document is longer than 2000 tokens, highlighting the difficulty for model to capture long-range contextual dependencies
- Readability and factuality metrics are less affected by the length of the input document. The model can generate fluent (DCRS) and factual (BARTScore) text with just a few tokens.

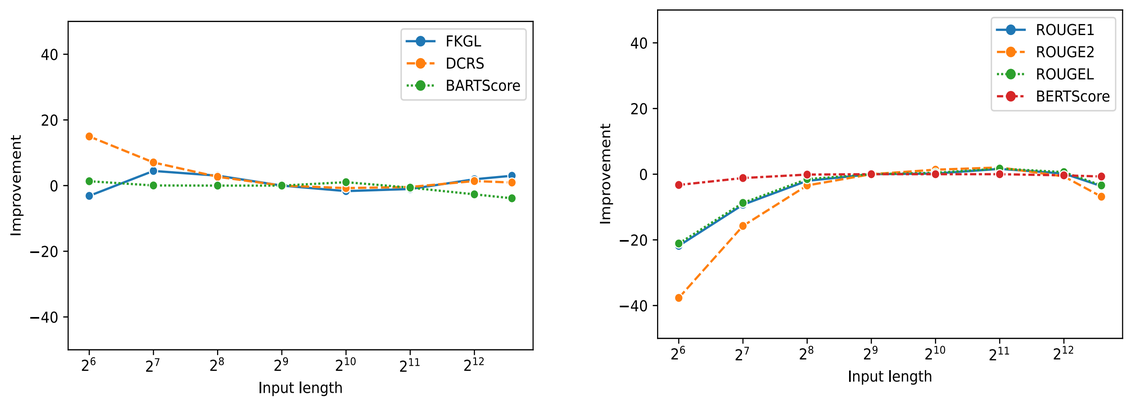


Figure 1: The impact of input sequence length on evaluation scores on the eLife development set. We use the input length of 512 (2^9) as the baseline and measure the relative improvement due to different input lengths.

Summary

We utilise GPT-3.5 to generate additional training data and pre-trained FLAN-T5 in generating a lay summary for scientific documents. Our results show that extra data generated from a generative model can boost the effectiveness of a summarisation model to a certain degree, especially in terms of relevance metrics.

FOR FURTHER INFORMATION

Mong Yuan Sim
e mongyuan.sim@student.adelaide.edu.au

Sarvnaz Karimi
e sarvnaz.karimi@csiro.au

REFERENCES

Mong Yuan Sim, Xiang Xiang, Maciej Rybinski and Sarvnaz Karimi. 2023. CSIRO Data61 Team at BioLaySumm Task 1: Lay Summarisation of Biomedical Research Articles Using Generative Models. In *Proceedings of the 22nd Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

ACKNOWLEDGEMENTS

This work is supported by The Commonwealth Scientific and Industrial Research Organisation (CSIRO) Precision Health Future Science Platform (FSP). Experiments were undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.