Improving Factuality in Clinical Abstractive Multi-Document Summarization through Guided Continued Pre-training

Ahmed Elhady[1,2] **Khaled Mostafa Elsayed**[2] **Eneko Agirre**[1] **Mikel Artetxe**[1,3]
[1]HiTZ Center, University of the Basque Country (UPV/EHU)
[2]University of Science and Technology, Zewail City [3]Reka AI
{ahmed.salemmohamed,e.agirre,mikel.artetxe}@ehu.eus
{kmelsayed}@zewailcity.edu.eg

Abstract

Factual accuracy is an important property of neural abstractive summarization models, especially in fact-critical domains such as the clinical literature. In this work, we introduce a guided continued pre-training stage for encoder-decoder models that improves their understanding of the factual attributes of documents, which is followed by supervised fine-tuning on summarization. Our approach extends the pre-training recipe of BART to incorporate 3 additional objectives based on PICO spans, which capture the population, intervention, comparison and outcomes related to a clinical study. Experiments on multi-document summarization in the clinical domain demonstrate that our approach is competitive with prior work, improving the quality and factuality of the summaries and achieving the best published results in factual accuracy on the MSLR task.

1 Introduction

Neural abstractive Multi-Document Summarization (MDS) is an active area in natural language processing. It requires comprehension of several input documents by resolving the shared and potentially redundant information among them, and the generation of fluent salient summaries (Ma et al., 2022; Nallapati et al., 2016; Chopra et al., 2016). While advances in sequence-to-sequence models have improved the fluency and cohesion in abstractive summarization, the generation of unfaithful or non-factual summaries is still a major issue (Li et al., 2022). This is critical in many settings like clinical document summarization, where factual accuracy is more valued than other summary qualities (Wallace et al., 2020).

An effective method for improving the factuality of summaries is to guide the model using additional guidance signals extracted from the input (Li et al., 2022). Models are then trained to either generate the signals prior to the summary or condition the generation of the signals by prepending them to the source text. Signals included entity chains (Narayan et al., 2021; Zhang et al., 2022), keywords and custom prompts (He et al., 2020), legal arguments (Elaraby and Litman, 2022), and a mixture of human-annotated attributes (Zhang et al., 2023). Dou et al. (2020) introduced a generalized framework for guided summarization (*GSum*) where they use a secondary encoder for the guidance signals, and combine it with the original document encoder in BART (Lewis et al., 2019). The decoder then uses both encoders' outputs for generation.

However, all these approaches operate at the supervised fine-tuning stage, and are thus constrained by the amount of labeled data available. In addition, they assume fine-tuning will inherently make the model pay higher attention to the factual attributes in input documents, which is neither guaranteed nor sufficient for generating faithful summaries (Wallace et al., 2020).

In this work, we propose a guided continued pre-training approach for improving the model's understanding of the factual constitution of the documents using guidance signals. Our method operates over unlabeled corpora from the target domain, and the model is then fine-tuned directly on the summarization task without additional inputs or attribute highlights. Our pre-training objective extends BART (Lewis et al., 2019) to help the model understand meta-information about the factual attributes of the documents based on PICO spans.

Our main contributions are: 1) we introduce a set of continued pre-training objectives addressing factual attributes; 2) we show that the proposed method outperforms baselines and prior methods improving both the summary quality and factuality, with ablation of the contribution of each objective to the performance gain; and 3) we show that the proposed technique improves the faithfulness of the summaries in few-shot and zero-shot settings. To the time of writing, our method ranks 1st in the

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 755–761

Cochrane task leaderboard.[1]

2 Proposed method

Our proposed approach consists of 3 steps: (i) pre-training a general-purpose encoder-decoder model (§2.1), (ii) specializing the model in the target domain using a continued pre-training stage (§2.2), and (iii) supervised fine-tuning of the domain-specialized model for MDS (§2.3).

2.1 General pre-training

As the first step, we pre-train a general purpose encoder-decoder model on general domain corpora. In our experiments, we do not run pre-training ourselves, and use BART-large instead.

2.2 Guided continued pre-training

In the second step, we specialize the model using documents from the target domain and a set of guided pre-training objectives that enrich the model's understanding of the factual constitution of the domain document, which is the main contribution of our work. To that end, we rely on PICO elements,[2] which we automatically extract using BioElectraPICO (Kanakarajan et al., 2021). We base this choice on the findings of previous work, which showed a direct correlation between these elements and the factual accuracy of clinical documents (Huang et al., 2006; Wallace et al., 2020; DeYoung et al., 2021). Our method combines the following objectives:

BART objective. The original text infilling and sentence permutation objectives from Lewis et al. (2019).

PICO infilling. A special case of text infilling, where we randomly select some PICO elements for masking. The masked span starts at the beginning of each PICO element and its length is sampled from $\min(\text{Poisson}(\lambda = 3), \text{len}(\text{PICO}))$.

PICO infilling with special masks. Equivalent to *PICO infilling*, but instead of replacing the selected spans with the general masking token <*mask*>, we replace them with special masks corresponding to their PICO annotation (*population*, *intervention*, *comparison* or *outcome*).

[1] https://leaderboard.allenai.org/mslr-cochrane/submissions/public

[2] PICO elements are text segments representing the Population, Intervention, Comparison, and Outcomes related to a clinical study or review.

Original Text To evaluate the acute and chronic physiotherapy effects of these two techniques, 14 cystic fibrosis patients underwent either twice daily autogenic drainage or Flutter treatment for 4 consecutive weeks in a randomized crossover design.

BART's Text Infilling To evaluate the acute and chronic <*mask*> these two techniques, 14 cystic fibrosis patients underwent either twice daily autogenic drainage or Flutter treatment for 4 <*mask*> in a randomized <*mask*>.

PICO Infilling To evaluate the <*mask*> physiotherapy effects of these two techniques, <mask> underwent either twice daily <mask> or Flutter treatment for 4 consecutive weeks in a randomized crossover design.

PICO Infilling with Special Masks To evaluate the <*intervention*> physiotherapy effects of these two techniques, <population> underwent either twice daily <*intervention*> or Flutter treatment for 4 consecutive weeks in a randomized crossover design.

Figure 1: An example of the three types of text infilling corruptions. PICO element spans are color-coded according to their type. For all types of infilling, masked span lengths are randomly sampled from a Poisson distribution ($\lambda = 3$).

Guided GSG. For a document $D = \{x_i\}_n$ consisting of n sentences, each with p_i PICO elements, we compute the following score for each sentence:

$$s_i = \text{ROUGE}_1(x_i, D \backslash \{x_i\}) + \frac{\text{num\_toks}(p_i)}{\text{num\_toks}(x_i)}$$

The first term is the Rouge-1 score between the sentence in question and the rest of the document with that sentence removed, and the second term is the proportion of tokens in the sentence that fall within a PICO span. Sentences are then sorted by their scores, and the top 30% are selected as the target. These target sentences are removed from the input, and the model is trained to reconstruct them. This is analogous to the Gap Sentence Generation (GSG) objective from Zhang et al. (2019), except that we add an additional term to the score to favor sentences with a high proportion of PICO tokens. Therefore, the selection of gap sentences is guided towards both factual and important sentences.

We apply the objectives above on 50% of source-target pairs per epoch, distributed as follows: 15% for PICO infilling, 15% for guided GSG, 10% for the PICO infilling with special masking, and 5% for each of BART's infilling ad sentence permutation objectives. No changes are applied to the remaining

50%. Figure 1 shows examples to illustrate the different types of text infilling.

2.3 MDS fine-tuning

In the third and last step, we fine-tune the model on abstractive MDS. We concatenate all input documents, truncating sequences longer than 2048 tokens, and train the model to predict the reference summary.

3 Experimental settings

Dataset. We experiment with AllenAI's Multi-document Summarization of Literature Review (MSLR) task. The task consists of two datasets: MS^2 (DeYoung et al., 2021) and Cochrane (Wallace et al., 2020). In both datasets, the inputs are abstracts of clinical papers, and the targets are either the abstract of their corresponding review paper in MS^2, or the author's conclusion in Cochrane. We choose this dataset because it provides a quantitative metric for evaluating factual consistency. Results are reported on the validation sets using the official task's evaluation script. Appendix A reports additional results on the test set.

Continued pre-training. We sample 360k non-empty English abstracts from the PubMed dataset on HuggingFace[3] on the same clinical domain as the MSLR datasets. This is the same number of input abstracts in the MS2 and Cochrane training sets combined. Pre-training is done for 50 epochs, using a learning rate of 3e-05. The maximum sequence length is set to 2,048 tokens. Batch size of 32 was used. We used the default setting of FairSeq (Ott et al., 2019) for the rest of hyperparameters.

Fine-tuning. In our setup, input abstracts are grouped by Review Id and concatenated to a maximum of 2048 tokens. To assess the generalization capabilities achieved by our suggested pre-training method, we experimented with full-shot fine-tuning, few-shot fine-tuning (using 10% of the training data), and zero-shot learning.[4] Fine-tuning is done for 20,000 and 5,000 total number of updates for full and few-shot settings, respectively, with a learning rate of 3e-05. We use a dropout of 0.1 in the full-shot setting, and 0.15 in the few-shot setting. Batch size was set to 16.

[3] https://huggingface.co/datasets/pubmed

[4] Zero-shot learning uses the pre-trained model without any supervised fine-tuning. This can potentially generate sensible summaries thanks to the guided GSG objective in continued pretraining.

Objective	R-L↑	R-1↑	R-2↑	ΔEI↓
No continued pretrain	15.79	23.31	6.09	37.72
GSG (Zhang et al., 2019)	18.09	29.42	7.12	37.61
Our method	**19.82**	**29.88**	**7.40**	**34.70**
- *BART obj only*	19.10	26.39	6.17	38.55
- *PICO masking only*	18.53	24.11	5.29	36.42
- *Guided GSG only*	19.13	28.25	6.78	36.95

Table 1: Full-shot results on Cochrane, using different objectives for continued pretraining.

Metrics. To measure the quality of the generated summary, we report the Rouge-(1/2/L) (Lin, 2004), which measures the token-based similarity between generated and reference summaries. For factual consistency, ΔEI (DeYoung et al., 2021) is used, which measures the Evidence Inference (DeYoung et al., 2020) consistency between input documents and both generated and reference summaries, and then calculates the Jensen-Shannon Distance (Menéndez et al., 1997) between them. The closer the ΔEI is to zero, the factually closer the generated summary is to the reference. We report the F-1 score for the ΔEI for our experiments results.

4 Results

4.1 Main results

We do continued pre-training over BART using the proposed objectives individually, and evaluate the resulting model after supervised fine-tuning on Cochrane. As shown in Table 1, doing continued pretraining without any of the PICO-based objectives does not enhance the factual accuracy of the generated summaries, although the general summary quality in terms of Rouge is considerably better (*BART obj only* vs. *no continued pretrain*). Text infilling with PICO masking, on the other hand, yields the most factually consistent results, reducing the ΔEI by 1.3% from the baseline. Results of GSG and guided GSG show that 1) both objectives improve the quality of the generated summaries, but not necessarily factuality, and 2) guiding the GSG objective helps improve the factual accuracy of the generated summaries without sacrificing quality gains. Combining all of our proposed objectives obtains the best results by a substantial margin.

	Cochrane				MS^2			
	R-L↑	**R-1↑**	**R-2↑**	**ΔEI↓**	**R-L↑**	**R-1↑**	**R-2↑**	**ΔEI↓**
BART (Obonyo et al., 2022)	16.43	22.48	6.00	38.23	10.17	13.18	1.31	42.53
BART† (Lewis et al., 2019)	15.79	23.31	6.09	37.72	11.46	13.99	1.89	41.80
Longformer (Wang et al., 2022)	17.60	23.90	6.60	33.20	19.60	26.40	8.00	41.20
ITTC-2 (Otmakhova et al., 2022)	18.40	24.60	6.90	**30.90**	-	-	-	-
LED-base-16k (Giorgi, 2022)	18.03	25.73	6.58	39.94	**20.60**	**27.50**	**9.20**	42.40
PuneICT (Tangsali et al., 2022)	17.30	24.70	5.50	37.90	14.40	20.60	3.50	35.60
GSum† (Dou et al., 2020)	19.70	**30.71**	**7.82**	39.52	15.23	22.89	4.41	34.62
Our Method	**19.82**	29.88	7.40	34.70	14.27	19.83	3.18	**27.38**

Table 2: Comparison with prior work. ↑ means higher is better, ↓ means lower is better. †Results of our runs.

		R-L↑	**R-1↑**	**R-2↑**	**ΔEI↓**
Few-shot	Our Method	**17.98** ±0.3	**23.95**±1.2	**4.41**±0.3	**33.02**±0.7
	GSum†(Dou et al., 2020)	14.12 ±0.2	21.78±0.3	3.67±0.1	34.25±1.4
Zero-shot	Our Method	**13.60**	**22.30**	**3.10**	**35.70**
	Gsum†(Dou et al., 2020)	12.50	21.60	2.90	37.60

Table 3: Few-shot and zero-shot results on Cochrane. In the few-shot setup, we use 10% of the data and perform 5 runs for each model with different random seeds, and report the average and standard deviation of the results. †Results of our runs.

4.2 Comparison with prior work

Table 2 shows the comparison of our method to prior work on the Cochrane and MS^2 datasets. We also compare our method to GSum as the generalized framework of guidance-based summarization systems typically used to improve factual consistency in the non-clinical domain. Generally, we achieve state-of-the-art results outperforming all methods in factual accuracy on both datasets, with the exception of ITTC-2 and Longformer on Cochrane, where both use larger-size models.

Cochrane. Results show that our method consistently improves over published methods that use BART (Obonyo et al., 2022; Tangsali et al., 2022). We outperform the Longformer-based ones (Wang et al., 2022; Giorgi, 2022) in Rouge scores as well. GSum achieves similar enhancements to ours in Rouge scores, yet no gain in ΔEI. We also observe tendency to repeat information and verbatim text segments in the generated summaries of GSum. Appendix B contains examples of this behavior.

MS^2. Compared to Cochrane, MS^2 consists of longer abstracts making the constraining maximum sequence length of our method more limiting. This gives advantages for techniques that accept long sequences such as Longformer (Wang et al., 2022; Giorgi, 2022). Despite that, our technique yields the best factual consistency results, reducing the ΔEI by 15% compared to the best-performing Longformer model.

4.3 Few- and zero-shot learning

Table 3 reports few-shot and zero-shot results on Cochrane. Zero-shot results suggest that the continued pre-training helps generate factually consistent summaries. Further improvements in quality and factual accuracy can be acquired by fine-tuning with as few samples as 10% of the data. Compared to full fine-tuning results in Table 2, few-shot achieves better ΔEI scores, suggesting the effectiveness of our method with limited resources. We also noticed that generated summaries of the full fine-tuning model tend to be shorter than those generated by the few-shot models. This can be explained by the variation in lengths of the reference summaries in the dataset and explains the enhanced factual consistency in the few-shot setup.

5 Conclusions

We propose a continued pre-training stage that combines several objectives designed to improve the understanding of clinical documents. The resulting model improves the factual accuracy of summaries, even in few- and zero-shot settings. With limited resources, our system based on BART-large out-

performs models of the same scale and achieves competitive results with larger ones. All in all, our work demonstrates that it is possible to adapt a general purpose model to a specific task and domain without any labeled data by continuing pre-training with a carefully designed objective.

Limitations

We base our hypothesis on the direct mapping between guidance signals and factual attributes of documents. However, the selection of PICO elements as the guidance signals in our method is based on prior work that assumed a correlation between them and factual accuracy. Other choices of attributes, such as evidence sentences, have not been explored.

Another limitation is that our base model, BART, has a maximum sequence length of only 2048, which results in many input sequences being truncated. Sequence-to-sequence models that accept longer sequences, such as Longformers (Beltagy et al., 2020), were not explored due to hardware limitations. Despite that, our continued pre-training method is generic and does not use any BART-specific features, thus it can be safely assumed it would improve results for other language model architectures. Finally, even if our approach improves the factual accuracy of the generated summaries, outputs need to be used with care as they may still contain incorrect information.

Acknowledgements

This work has been partially supported by the Basque Government (Research group funding IT-1805-22). Ahmed Elhady holds a PhD grant supported by the Basque Government (IKER-GAITU project).

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS2: multi-document summarization of medical studies. *CoRR*, abs/2104.06486.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *CoRR*, abs/2010.08014.

Mohamed Elaraby and Diane Litman. 2022. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

John Giorgi. 2022. Fine-tuned version of allenai/led-base-16384 on the cochrane dataset. `https://huggingface.co/allenai/led-base-16384-cochrane`. Accessed: 2023-11-15.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Fatema Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *CoRR*, abs/2012.04281.

Xiaoli Huang, Jimmy J. Lin, and Dina Demner-Fushman. 2006. Evaluation of pico as a knowledge representation for clinical questions. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, pages 359–63.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Comput. Surv.*, 55(5).

M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Ramesh Nallapati, Bing Xiang, and Bowen Zhou. 2016. Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, and Ryan T. McDonald. 2021. Planning with entity chains for abstractive summarization. *CoRR*, abs/2104.07606.

Ishmael Obonyo, Silvia Casola, and Horacio Saggion. 2022. Exploring the limits of a base BART for multi-document summarization in the medical domain. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 193–198, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yulia Otmakhova, Thinh Hung Truong, Timothy Baldwin, Trevor Cohn, Karin Verspoor, and Jey Han Lau. 2022. LED down the rabbit hole: exploring the potential of global attention for biomedical multi-document summarisation. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 181–187, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Rahul Tangsali, Aditya Jagdish Vyawahare, Aditya Vyankatesh Mandke, Onkar Rupesh Litake, and Dipali Dattatray Kadam. 2022. Abstractive approaches to multidocument summarization of medical literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 199–203, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain James Marshall. 2020. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *CoRR*, abs/2008.11293.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Cochrane				
	R-L↑	**R-1↑**	**R-2↑**	**ΔEI↓**
PuneICT	19.69	26.22	5.74	**30.11**
ITTC-2	18.37	24.64	**6.92**	30.89
LED-base-16k	18.03	25.73	6.58	39.94
Our Method	**19.97**	**28.06**	6.67	35.35
MS^2				
	R-L↑	**R-1↑**	**R-2↑**	**ΔEI↓**
PuneICT	14.39	20.60	3.50	35.58
LED-base-16k	**20.60**	**27.48**	**9.20**	42.36
Our Method	14.80	19.81	3.40	**28.35**

Table 4: Results on the test set compared to other published results.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022. Improving the faithfulness of abstractive summarization via entity coverage control.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. MACSum: Controllable Summarization with Mixed Attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

A Test results

Table 4 shows results of our model compared to published methods on the blind test set. We show a consistent performance on the test set similar to that of the validation ones.

B Example summaries

Table 5 shows two examples of generated summaries of our method using BART-Large and GSum. We noticed a tendency in GSum's summaries to repeat ideas and/or verbatim PICO elements guidance signals. For example, GSum's summary in the first sample repeats the PICO element *Congential Heart Disease*, and repeats the idea of *postoperative shivering* in the second sample. This repetition suggests reduced factual accuracy despite an increase in the ROUGE scores.

Input Abstracts	Summary	
Atrial fibrillation (AF) is a common arrhythmia in adults with congenital heart disease (CHD). Long-term antiarrhythmic therapy (AAT) in these patients has significant shortcomings. The safety and efficacy of pulmonary vein antrum isolation (PVAI) for the treatment of AF in CHD is presently unknown . We hypothesized that PVAI for AF in patients with CHD is effective and safe. We reviewed a prospective cohort of 4315 patients undergoing PVAI for drug refractory AF at a single institution and identified 36 consecutive patients with CHD (single ventricle physiology , tetralogy of Fallot, coarctation of the aorta, ventricular septal defects , atrial septal defects (ASD) ...	Our Method	Pulmonary vein antrum isolation (PVAI) is a safe and effective treatment for AF in patients with CHD . There is no evidence of a difference in the risk of adverse events between the two groups , but there is a trend towards a higher risk of pulmonary vein thromboembolism in the PVAi group compared with the non-PVAI group.
	GSum	PVAI is an effective treatment for AF in patients with congenital heart disease and non-Congenital Heart Disease. However, there is insufficient evidence to support its use in patients with non Congenital Heart Disease. Further well-designed randomised controlled trials are needed to determine the role of PVAI in the treatment of AF in CHD.
Intrathecal adjutants can be used for regional anesthesia (RA) in cesarean section to improve its quality in terms of time and complications. Some previous studies focused on the effects of adding sufentanil and /or midazolam to bupivacaine and compared each with using bupivacaine alone. However , there has been no study to assess the effects of using sufentanil and midazolam in combination with bupivacaine. The aim of this study was to evaluate and compare properties (time of achievement/recovery of sensory/motor blocks;and time to request opium), complications (nausea, vomiting, shivering and hypotension), and neonatal first minute Apgar score with and without the addition of midazolam (M) or sufentanil (S) to bupivacaine (B) through intrathecal injection for spinal anesthesia , after the cesarean section...	Our Method	The addition of midazolam or sufentanil to bupivacaine in addition to local anaesthetics does not appear to have any significant effect on the time to achieve spinal block, time to request opium, or time to request motor block recovery. However, the addition of midazolam or sufentanil to bupivacaine is associated with a significant increase in the risk of postoperative nausea and vomiting. Further research is needed to determine whether the addition of midazolam or sufentanil to bupivacaine in addition to local anaesthetics reduces the risk of postoperative adverse effects.
	GSum	There was no evidence of a difference in the incidence of postoperative shivering . Conclusions There is no evidence to support the use of sufentanil or midazolam as an adjunct to bupivacaine in caesarean section for the prevention of shivering or postoperative vomiting.

Table 5: Example of generated summaries of our method and our runs of GSum. Beam-search width is set to 5 in both experiments.