Multi-document Summarization in Medical Literature using PICO-Masking Approach

Anonymous ACL submission

Abstract

Multi-document summarization is essential for capturing key information from vast medical literatures. The dataset of this domain typically comprises a triple of a background, documents and a summary where background describes clinical research question or topics shared by related documents. To summarize based on a background while accommodating multiple documents, existing approaches typically reduce text units through truncation disregarding potential summary-relevant information. Others perform extract-then-generate approaches at document-level or sentence-level which could struggle to capture the relevant evidence since document-level extraction is excessively broad and sentence-level extraction is overly granular and noisy. To address the aforementioned problems, we combine two extraction levels and propose to frame the problem as query-focused summarization where background represents a query. Specifically, we decompose the problem into two stages 1) *relevant evidence extraction* (i.e. finding relevant evidence within a set of relevant documents with regards to the shared background) 2) *summary generation* (i.e. generating summaries based on the relevant evidence). To represent background as a query, we introduce a PICO-masking approach to mask the given background and consider it as a *proxy query* for our extraction model. In particular, PICO-masking masks elements that are mnemonic for the important parts of a well-built clinical question. This enforces extraction model to understand the context in order to identify the evidence from documents that belong to the masked background, hence help locate relevant evidence before generating a summary. Results show that our approach achieves state-of-the-art performance on MS2 dataset despite having multiple stages.

1 Introduction

Multi-document summarization is essential for capturing key information from several documents. It has been applied to many domains such as news summarization (Fabbri et al., 2019b), Wikipedia articles (Liu et al., 2018),and scientific articles (Lu et al., 2020). In medical domain, significant research efforts have been directed towards developing effective summarization approaches for handling extensive medical documents. Specifically, the dataset of this domain comprises a triple of a background, documents and a summary where background describes clinical research question or topics shared by related documents. To summarize based on a background while accommodating multiple documents, we identify two typical approaches to reduce text units 1) truncation and 2) extract-then-generate approach where truncation disregards potential summary-relevant information that may be located at particular location of the documents (DeYoung et al., 2021; Tangsali et al., 2022; Wang et al., 2022) and extract-then-generate approaches at document-level (Moro et al., 2022) or sentence-level (Shinde et al., 2022) which could struggle to capture the relevant evidence since document-level extraction is excessively broad and sentence-level extraction is overly granular and noisy (Xu and Lapata, 2020).

To address the aforementioned problems, we combine two extraction levels and propose to frame the problem as query-focused summarization where background represents a query. Specifically, we decompose the problem into two stages 1) *relevant evidence extraction* (i.e. finding relevant evidence within a set of relevant documents with regards to the shared background) 2) *summary generation* (i.e. generating summaries based on the relevant evidence). To represent background as a query, we introduce a PICO-masking approach to mask the given background and consider it as a *proxy query* for our extraction model. In particular, PICO-masking marks elements that are mnemonic for the important parts of a well-built clinical question. This enforces extraction model to understand the context in order to identify the evidence from documents that belong to the masked background, hence help locate relevant evidence before generating a summary. In summary, our approach applies no input truncation, while enabling relevant evidence allocation.

Our contributions in this work are threefold: we are the first to frame multi-document summarization in medical literature as query-focused summarization; we discover a specific masking approach for this domain; we provide experimental results and show that our proposed approach achieve state-of-the-art result on MS2 dataset.

2 Related Work

We reviewed related areas of research: (1) multi-document summarization in medical literatures, (2) query-focused summarization, (3) masking techniques.

2.1 Multi-document Summarization in Medical Literature

Substantial progress has been achieved in multi-document summarization (Fabbri et al., 2019a; Liu et al., 2018; Lu et al., 2020). In the domain of medical literature which aims to create a summary from multiple documents based on a shared background is comparatively less prevalent when compared to multi-document summarization in other domains. Similar to other domains, medical literature also faces problem with long input text. To reduce text units, two approaches have been explored. First, truncation has been applied to concatenated documents and background before fine-tuning long-range neural models (DeYoung et al., 2021; Tangsali et al., 2022; Wang et al., 2022). Second, extraction is applied to extract relevant documents or sentences before generating a summary (Shinde et al., 2022; Moro et al., 2022). However, truncation disregards potential summary-relevant information that may be located at particular location of the documents and extracting relevant information at document-level or sentence-level may struggle to generate accurate summary since document-level extraction is excessively broad and sentence-level extraction is overly granular and noisy which is a matter of particular concern, as the discussed subject pertains to the medical domain.

In this work, we follow extract-then-generate approach. Specifically, we decompose the problem into two stages 1) *relevant evidence extraction* (i.e. finding relevant evidence within a set of relevant documents with regards to the shared background) 2) *summary generation* (i.e. generating summaries based on the relevant evidence). However, in contrast to the work of Shinde et al. that extracts relevant information at document-level and the work of Moro et al. that extracts relevant information at sentence-level, our relevant evidence extraction model combines the two by first extract relevant documents and from those documents relevant evidence is extracted.

2.2 Query-focused Summarization

Query-focused summarization (QFS) is known as an important extension for summarization. It focuses on generating concise summaries tailored to a specific query. The dataset in this domain typically comprises a triple of document, query and a summary. Early efforts in this domain primarily revolved around unsupervised extractive approaches (Wan et al., 2007; Litvak and Vanetik, 2017) due to limited availability of training data (Dang, 2005).

Recent advancements have leveraged the relationship between query-focused summarization and the more data-abundant task of question answering for extractive summarization (Egonmwan et al., 2019), keyword mapping (He et al., 2020), document reranking within a retrieval pipeline (Su et al., 2020), and abstractive summarization (Su et al., 2021; Baumel et al., 2018; Yujia et al., 2020; Xu and Lapata, 2020)

Given its success and similarity in generating summary tailored to a specific need, we see the opportunity in framing our problem as QFS. Note that in our case, a query is absent and our generated summary is tailored to a shared background.

2.3 Masking Techniques

Masking has been widely used in natural language processing tasks, contributing to the success of various models especially in the context of pre-training and fine-tuning transformers. This includes masked language modelling (Devlin et al., 2018; Liu et al., 2019) (i.e. masking input tokens at random allowing the model to learn contextualized word representations), sentence completion (i.e. predicting the masked fraction of a sentence is masked - Cloze tests)(Taylor, 1953), question-answering (i.e. masking relevant portion of input text allowing the model to predict the missing information) (Jun et al., 2022), named entity recognition (i.e. replacing name entities allowing the model to recognize and classify the entities) (Sonkar et al., 2022), domain adaptation (i.e. injecting domain specific knowledge emphasizing relevant vocabulary) (Gu et al., 2020; Lamproudis et al., 2021), etc. In short, masking is a promising approach to enhance understanding of context and promote context comprehension.

In the area of QFS, Xu and Lapata proposed an approach to transform generic summarization datasets into query-focused training data through masking. Specifically, inspired by Cloze task (Taylor, 1953), Xu and Lapata inproduced *Unified Masked Representation (UMR)* to convert summary to proxy query used during training. Specifically, document sentences are parsed to Open Information Extraction (Open IE; (Stanovsky et al., 2018)) to obtain a set of a propositions consisting of verbs and their arguments. Then according to certain budget constrain, the arguments are replaced with `[MASK]` tokens.

In contrast, instead of argument masking, we propose PICO-masking specifically for our medical literature summarization. In particular, PICO-masking masks elements that are mnemonic for the important parts of a well-built clinical question. This enforces extraction model to understand the context in order to identify the evidence from documents that belong to the masked background.

3 Method

We propose to frame the problem as query-focused summarization. Let $\{D, S\}$ denote single document summarization dataset D denote a documents and S is a summary. In query-focused summarization, it additionally provides a query Q for summary generation, $\{(D, Q, S)\}$.

On the other hand, in the area of multi-document summarization in medical literatures, instead of a query Q, it provides a background B that describes clinical research question or topics shared by documents for summary generation, $\{(D, B, S)\}$. Additionally, in contrast to single document summarization where D denote a document, here D denote a set of documents, $D = \{d_1, d_2, ..., d_M\}$.

Specifically, we decompose the problem into two subtasks; namely 1) *relevant evidence extraction* and 2) *summary generation*. Note that our relevant evidence extraction is further decomposed into *candidate document extraction* (i.e. document-level extraction) and *candidate sentence extraction* (i.e. sentence-level extraction) whose aim here is to allocate relevant evidence from identifying relevant documents to relevant sentences. Here, **c**andidate **d**ocument extraction model $c_{d,\theta}(\hat{D}|B;\theta)$ extracts relevant documents $\hat{D}$ to background B within a set of documents D and **c**andidate **s**entence extraction model $c_{s,\phi}(\hat{C}|\hat{D}, \hat{B};\phi)$ then extracts relevant sentences $\hat{C}$ to background B within a set of relevant document $\hat{D}$. Note that $\hat{B}$ denote a masked background which serves as a *proxy query* to train our candidate sentence extraction model. Then, $g_\varphi(S|\hat{C}, B;\varphi)$ generates summary S conditioned on evidence provided by the relevant evidence extraction and the background itself.

To convert background B to serve as *proxy query*, we were inspired by *Unified Masked Representation (UMR)* proposed by (Xu and Lapata, 2021). Here, we also assume that answers to the query are located within the sentences in the set of relevant documents $\hat{D}$. Here we refer sentences that contain answers as relevant sentences. As it is uncertain which sentences contain the answers, we presume their relevance by assuming a high ROUGE score against the query. Hence, we employ ROUGE as our distant supervision signal to train our candidate sentence extraction model to extract relevant sentences from a set of relevant documents and a background. The most relevant sentences then serve as an input to the summary generation model along with the background.

3.1 Relevant Evidence Extraction

Our relevant evidence extraction comprises two parts which are 1) candidate document extraction and 2) candidate sentence extraction. Specifically, candidate document extraction involves identifying relevant documents, while candidate sentence extraction extracts relevant sentences. Next we explain each part in details.

3.1.1 Candidate Document Extraction

We extract candidate documents using Dense Passage Retrieval (DPR) (Karpukhin et al., 2020). Here DPR is selected due to its ability to provide a deeper semantic understanding of documents allows for more accurate and contextually relevant selections. Here top-6 documents are extracted (Moro et al., 2022).

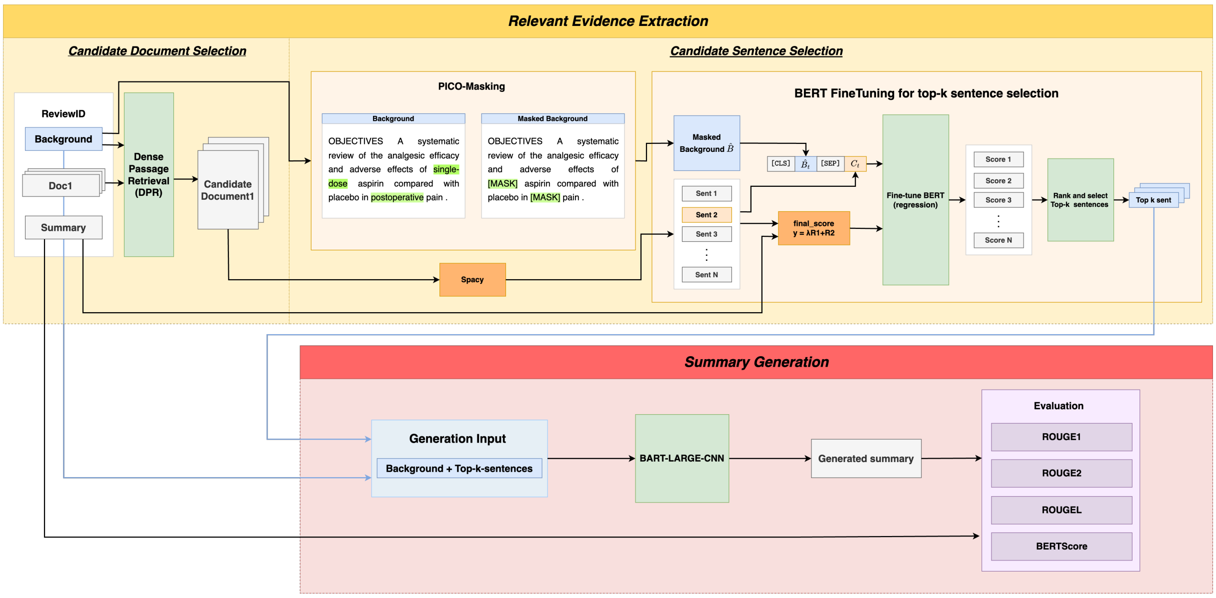

Figure 1: Overview of our framework

3.1.2 Candidate Sentence Extraction

Here we were inspired by *Unified Masked Representation (UMR)* proposed by (Xu and Lapata, 2021). Specifically, (Xu and Lapata, 2021) renders query from reference summary by replacing a small fraction of query with `[MASK]` to represent missing information that can be found in the document. Similarly, we also covers a small fraction, but of the background.

To identify which fractions to replace, we introduce PICO-masking approach. In particular, PICO is a framework that describes several essential components of the central question in a clinical trial, including Populations (e.g. diabetics), Interventions (e.g. animal insulin), Comparators (e.g. human insulin), and Outcomes (e.g. glycaemic control) (Huang et al., 2006). It aids in constructing the search strategy by locating the concepts necessary in medical documents that can address the posed question. By masking PICO elements, we hypothesize that it would enforce our extraction model to understand the context in order to identify the evidence from documents that belong to the masked background, hence help extract relevant sentences.

To perform PICO-masking, we employ Bio-Electra model (Kanakarajan et al., 2021) to identify PICO-elements in selected document sentences. Here Bio-Electra model, a biomedical domain-specific language model, is selected due to its high performance in discerning PICO elements within a document. Specifically, PICO elements found $P = \left\{p_1, p_2, ..., p_{|P|}\right\}$ are partially replaced with `[MASK]`. Here the masking percentage is kept at 15% (See Table 7 for our selected masking percentage justification).

To extract relevant sentences, we employ a pre-trained BERT model (Devlin et al., 2018) to regressively rank document sentences based on relevant score. Specifically, we concatenate masked background with document sentence "`[CLS]` $\hat{B}_t$ `[SEP]` C_t `[SEP]`" where $\hat{B}_t$ denote a sequence of tokens of the masked background and C_t denote a sequence of tokens in document sentence. Given the input, we train our BERT model with the objective to minimize the mean-square error loss to regressitvely predict the relevant score.

$$L(\phi) = \frac{1}{|D|} \sum_{(\hat{B},C)\sim D} [(y - \hat{y}(\hat{B}, C))^2] \quad (1)$$

where $\hat{B}, C$ is a background-document sentence pair and y is the ROUGE training signal which is the F1 interpolation of ROUGE-2 and ROUGE-1 defined as:

$$y = R_2(\hat{B}, C) + \lambda \cdot R_1(\hat{B}, C) \quad (2)$$

where λ is set to 0.15 following the optimization of (Xu and Lapata, 2021). The highest ranked sentences are extracted and sent to our summary generation model.

Due to highly skewed score distribution of our training document sentences with over 85% of sentences scoring below 0.05, it leads to model over-

Statistics	Training	Validation	Test
Total Sample Count	14188	2021	400
Missing Background	210	38	0
Missing Target	42	0	0
Samples After Clean-up	13978	1983	400
Dropped ReviewIDs	210	38	0
Avg Tokens in Background	73.46	69.85	74.10
Avg Tokens in Target	61.26	60.89	59.57
Avg Tokens in Abstract	301.55	299.97	300.97

Table 1: Dataset statistics

fitting towards less relevant sentences. To overcome this, a low score sampling technique is applied. Specifically, pairs that yield less than 0.05 were removed. As the result, this promotes a more balanced generalizable training process. This adjustment not only aids in preventing model bias but also enhances computational efficiency, leading to a more robust model performance. Note that top-3 sentences are extracted due to its highest recall ROUGE-2 score against the summary. (See Table 7 for further details).

3.2 Summary Generation

To generate summary based on a shared background, we prepend the background to the relevant sentences. Specifically, we perform fine-tuning on the pretrained model. Given the input background and relevant sentences, the objective is to minimize the negative log-likelihood of generating output summary $S = \left\{s_1, s_2, ..., s_{|S|}\right\}$.

$$L(\varphi) = \sum_{i}^{|S|} \log P(s_i|\hat{C}, B, s_1, ..., s_{i-1}) \quad (3)$$

4 Experiment

4.1 Dataset

We perform experiments on the MS2 dataset for multi-document summarization in medical literature domain (DeYoung et al., 2021). It consists of 470K documents , 20K background and 20K summaries where documents consist of research papers, clinical trials and clinical reviews while background describes describes clinical research question or topics shared by related document and a summary encapsulates the overall findings. Due to the absence of background and target in some samples, those are disgarded which results in 14K training, 2k validation and 400 testing samples. The dataset statistics is shown in Table 1.

4.2 Experimental setting

Here we describe the experimental setting for each of the components of our work, namely the relevant evidence extraction which comprises candidate document extraction and candidate sentence extraction, and summary generation.

As for candidate document extraction, our implementation is based on the work of Moro et al.. Note that no training was performed at this stage.

On the other hand, for candidate sentence extraction, we performed our experiment on bert-base-uncased. Here all input was truncated to 512 tokens. For the fine-tuning, the learning rate is set to $1 \times 10^{-}3$ and the model was trained for 5 epochs at batch size 192. Additionally, we adopted Adam as our optimizer with weight_decay of 0.01 hyper-parameters. Note that we parsed our document inputs to spacy (Honnibal and Montani, 2017) to obtain document sentences and PICO elements were masked using Bio-Electra model (Kanakarajan et al., 2021). All input tokens are truncated to 512 tokens. Here, model identify each token into 4 different class i.e "I-Population","I-Intervention","I-Outcome" and "I-Others".

Last, for summary generation, bart-large-cnn was employed. Here all input was truncated to 1024 tokens and output is set to min and max of 32 and 256 tokens respectively. For the fine-tuning, the learning rate is set to $1 \times 10^{-}3$ and the model was trained for 3 epochs at batch size 4 with the min and max output lengths of 32 and 256 respectively. Additionally, we adopt Adam as our optimizer with default hyper-parameters. At inference time, beamsize of 4 is selected with the min and max output lengths are kept the same as fine-tuning. Note that all our language models were taken from HuggingFace.

As for the evaluation metric, following previous works, ROUGE scores (Lin, 2004) including ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L), and BERTScore (Zhang et al., 2020) were selected.

5 Results

Our experiments evaluate work against previous work by comparing the generated summary against its reference. Specifically, we calculate f1 ROUGE scores including ROUGE-1, ROUGE-2 and ROUGE-L of our generated sentences against the reference summary. Table 2 shows that our model outperformed other models on across f1

	R-1	R-2	R-L
MS^2-LED	26.89	8.91	20.32
MS^2-BART	27.56	9.40	20.80
DAMEN	28.95	9.72	21.80
Ext-Abs	26.22	5.74	19.69
BART-LARGE	21.39	3.49	14.49
Distill -BART-cnn-12-6	20.82	2.98	13.77
LED-base-16k	27.5	9.2	20.6
Long-T5- Pubmed	12.00	1.33	9.61
Ours	**32.89**	**10.79**	**21.85**

Table 2: Evaluation result on MS2 dataset, We compare the our results against previous work in terms of f1 ROUGE scores on testing set. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L recall respectively.

ROUGE scores including ROUGE-1, ROUGE-2 and ROUGE-L.

5.1 Ablation Study

We conducted ablation study to verify the effectiveness of our proposed PICO-masking and the choice of masking percentage in our work. In addition, we also present our justification on our top-3 sentence selection. Specifically, we compare our PICO-masking against various masking approaches including Random, NOUN, BM25 and TF-IDF (See Appendix A for implementation details).

5.1.1 Effectiveness of PICO-Masking

To verify the effectiveness of our proposed PICO-masking, we evaluate its effect in both of components of our work namely relevant evidence extraction and summary generation.

Relevant evidence extraction - we evaluate the result on extractive summarization metrics. In particular, we calculated ROUGE recall scores including ROUGE-1, ROUGE-2 and ROUGE-L of our extracted sentences against the reference summary. We present result in Table 3. The results show that PICO-masking outperforms other masking approaches followed by Noun, TF-IDF, while BM25 and Random are the lowest performers. Specifically, PICO outperforms Random and BM25 by 0.2 points on ROUGE-1 and ROUGE-L and 1 point on ROUGE-L. Note that NOUN yielded competitive results.

Summary generation - we evaluate the result on abstractive summarization metrics. In particular, we calculated ROUGE f1 scores including ROUGE-1, ROUGE-2 and ROUGE-L of our generated summary against the reference summary. We present result in Table 4. The results show that PICO-masking outperforms other masking approaches followed by Noun, TF-IDF, Random and BM25. Specifically, PICO outperforms other approaches by at least 1 ROUGE-1 scores. Note that Noun masking yielded a competitive results with PINO on ROUGE-2 and ROUGE-L.

	R-1	R-2	R-L
Random	44.62	12.12	28.15
BM25	44.62	12.12	28.15
Noun	44.83	12.99	28.60
TF-IDF	44.76	12.95	28.44
Ours (PICO)	**44.83**	**13.16**	**28.66**

Table 3: **Relevant evidence extraction performance** of PICO-masking against Random, BM25, Noun and TF-IDF at 15% masking percentage in recall ROUGE scores on testing dataset. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L recall respectively.

	R-1	R-2	R-L
Random	31.31	9.42	20.85
BM25	30.52	9.22	20.50
Noun	31.93	10.35	21.75
TF-IDF	31.61	9.68	20.96
Ours (PICO)	**32.89**	**10.79**	**21.85**

Table 4: **Summary generation performance** of PICO-masking against Random, BM25, Noun and TF-IDF at 15% masking percentage on f1 ROUGE scores on testing dataset. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L f1 respectively.

5.1.2 Effectiveness of Masking Percentage

To verify the effectiveness of the percentage of PICO-masking, we evaluate its effect in both of components of our work namely relevant evidence extraction and summary generation.

Relevant evidence extraction - we evaluate the result on extractive summarization metrics. In particular, we calculated ROUGE recall scores including ROUGE-1, ROUGE-2 and ROUGE-L of our extracted sentences against the reference summary. We present result in Table 5. The results show that 15% was the best performer followed by 30% and 45%. Hence, the trend of decreasing in generation performance as masking percentage increases is observed.

Summary generation - we evaluate the result on abstractive summarization metrics. In partic-

	R-1	R-2	R-L
15%	**44.83**	**13.16**	**28.66**
30%	42.67	12.87	27.96
45%	40.67	11.97	27.67

Table 5: **Relevant evidence extraction performance** of PICO-masking against 15%, 30% and 45% masking percentage on recall ROUGE scores on testing dataset. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L recall respectively.

ular, we calculated ROUGE f1 scores including ROUGE-1, ROUGE-2 and ROUGE-L of our generated summary against the reference summary. We present result in Table 6. The results show that 15% was the best performer followed by 30% and 45%. Hence, the trend of decreasing in generation performance as masking percentage increases is observed. (See Appendix B for further details)

	R-1	R-2	R-L
15%	**32.89**	**10.79**	**21.85**
30%	32.03	10.46	21.22
45%	31.85	10.25	21.01

Table 6: **Summary generation performance** of PICO-masking against 15%, 30% and 45% masking percentage on f1 ROUGE scores on testing dataset. R-1, R-2 and R-L are ROUGE-1, ROUGE-2 and ROUGE-L f1 respectively.

5.1.3 Effectiveness of Top-k Sentence Selection

To justify our top-k sentence selection, we evaluate its effect on our work namely on the relevant evidence extraction. We present result in Table 7. The results show that selecting top 3 sentences yielded highest ROUGE-2 recall score. Note that ROUGE-2 recall score increases before starts to decrease at top-3. This trend can be observed prominently our PICO-masking.

	R-2@1	R-2@2	R-2@3	R-2@4	R-2@5
Random	12.12	12.12	12.12	12.12	12.12
BM25	12.12	12.12	12.12	12.12	12.12
Noun	12.99	12.99	12.95	12.95	12.95
TF-IDF	12.95	12.95	12.95	12.95	12.95
PICO	13.04	13.14	13.16	13.04	13.11
Average	12.64	12.66	12.67	12.64	12.66

Table 7: Relevant evidence extraction performance R-2@k is ROUGE-2 recall against top k sentences

6 Discussion and Analysis

From our results, we can observe that PICO-masking outperforms other masking approaches, followed by Noun, TF-IDF, BM25 and Random. Here we further discuss the possible reasons behind it.

6.1 Masked words

To better understand the masking effect on our results, we obtain top-10 most frequently masked tokens of each masking approaches shown in Table 8. From the table, it is observed that words masked by PICO-masking are all medical related terms, followed by Noun and TF-IDF. On the other hand, words masked by BM25 and Random are non-medical related. This is no surprise due to the nature of each masking approach. For instance, TF-IDF and BM25 are frequency based, while Noun masks all the nouns present in the document and Random lacks specificity in word selection. The demonstration of the results emphasize that PICO-masking enforces the model to identify relevant evidence in the document (See Table 3), hence enable more effective summarizaiton (See Table 4).

Top-10	Random	BM25	Noun	TF-IDF	PICO
1	recurrences	Background	aim	systematic	patients
2	placed	Objectives	Purpose	Background	cancer
3	effectiveness	summarise	evidence	Objectivities	interventions
4	infiltrated	importance	review	increasing	efficacy
5	observational	systematic	duration	review	allergic
6	stimulation	Although	behavioural	optimal	muscle
7	management	PURPOSE	usefulness	summarise	clinical
8	caesarean	relatively	deficiency	study	therapy
9	mobilization	majority	patients	Adoption	antibiotic
10	demonstrated	subjective	thromboembolism	Malaria	preconditioning

Table 8: Top-10 most frequently masked token across different masking approaches

6.2 Selected Top-k Sentences

Next, we obtain commonly extracted sentences among PICO-masking and other masking approaches. The results demonstrate that PICO and Noun masking extracted most common sentences from the relevant documents, followed by TF-IDF, BM25 and Random. This is no surprise because Noun masking masks medical related terms than TF-IDF, BM25 and Random (as shown in Table 8). This emphasizes that masking medical related terms helps model identify relevant information in the document, hence generate effective summaries.

Masking approach	15%	30%	45%
PICO vs Noun	87	96	55
PICO vs TF-IDF	64	78	62
PICO vs BM25	58	67	40
PICO vs Random	52	60	51

Table 9: Comparison between extracted Top-3 sentences by different masking approaces. For instance, PICO vs Noun means how many extracted top-3 sentences are common among the two approaches.

7 Conclusion

In this work, we propose to frame a multi-document summarization in medical literatures as query-focused summarization which comprises of relevant evidence extraction and summary generation models. Specifically, our relevant evidence extraction is further decomposed to candidate document extraction (i.e. document-level extraction) and candidate sentence extraction (i.e. sentence-level extraction). Additionally, we also introduce PICO-masking approach as a way to represent background as a query. The results on MS2 dataset show that by framing the problem as query-focused summarization using PICO-masking, our proposed model outperformed state-of-the-art. Additionally, we also present extensive study of the effectiveness of our PICO-masking compared to other masking approaches (i.e. Random, BM25, Noun and TF-IDF) and our choice of masking percentage. The results show that framing the problem as query-focused summarization using PICO-masking is promising results.

Limitations

The limitations of this study are that this study only focused on the MS2 dataset and though our PICO-masking focuses medical related terms which we hypothesized to be beneficial for medical literature multi-document summarization, we only compared ours to random, frequency based masking (i.e. BM25, TF-IDF) and POS based masking (i.e. Noun), it is interesting to see whether attention-based masking would bring substantial benefit to the learning of our model. Last, different components of our model are independently train, hence it is interesting to explore an end-to-end training.

Ethical Considerations

The advancement in the development of complex neural network structures and the widespread availability of pre-trained language models have brought about substantial enhancements in the task of summarizing multiple documents. This task is particularly important in high-impact domains, especially in the medical field. Systematic literature reviews play a essential role in supporting the medical and scientific community. As a result, there is a need for robust assurances regarding the accuracy of the generated summaries. Existing state-of-the-art natural language processing (NLP) solutions fall short in providing such assurances, leading us to conclude that our proposed solution, like its predecessors, is not yet prepared for deployment. Further research is necessary to investigate more effective evaluation metrics for text summarization, and there is still a requirement for comprehensive accuracy assessments by medical professionals on a large scale. Additionally, if the proposed method is to be utilized with sensitive data like medical patient records, it must incorporate privacy-preserving policies.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable suggestion and feedback.

References

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies.

Elozino Egonmwan, Vittorio Castelli, and Md Arafat Sultan. 2019. Cross-task knowledge transfer for query-based text summarization. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 72–77.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019a. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019b. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020. Train no evil: Selective masking for task-guided pre-training. *arXiv preprint arXiv:2004.09733*.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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.

Changwook Jun, Hansol Jang, Myoseop Sim, Hyun Kim, Jooyoung Choi, Kyungkoo Min, and Kyunghoon Bae. 2022. Anna: enhanced language representation for question answering. *arXiv preprint arXiv:2203.14507*.

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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2021. Developing a clinical language model for swedish: Continued pretraining of generic bert with in-domain data. In *International Conference Recent Advances in Natural Language Processing (RANLP'21), online, September 1-3, 2021*, pages 790–797. INCOMA Ltd.

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.

Marina Litvak and Natalia Vanetik. 2017. Query-based summarization using mdl principle. In *Proceedings of the multiling 2017 workshop on summarization and summary evaluation across source types and genres*, pages 22–31.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv:2010.14235*.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 180–189, Dublin, Ireland. Association for Computational Linguistics.

Kartik Shinde, Trinita Roy, and Tirthankar Ghosal. 2022. An extractive-abstractive approach for multi-document summarization of scientific articles for literature review. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 204–209, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Shashank Sonkar, Zichao Wang, and Richard G Baraniuk. 2022. Maner: Mask augmented named entity recognition for extreme low-resource languages. *arXiv preprint arXiv:2212.09723*.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. 2020. Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management. *arXiv preprint arXiv:2005.03975*.

Dan Su, Tiezheng Yu, and Pascale Fung. 2021. Improve query focused abstractive summarization by incorporating answer relevance. *arXiv preprint arXiv:2105.12969*.

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.

Wilson L Taylor. 1953. “cloze procedure”: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI*, volume 7, pages 2903–2908.

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.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645, Online. Association for Computational Linguistics.

Yumo Xu and Mirella Lapata. 2021. Generating query focused summaries from query-free resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6096–6109, Online. Association for Computational Linguistics.

Xie Yujia, Zhou Tianyi, Yi Mao, and Chen Weizhu. 2020. Conditional self-attention for query-based summarization. *arXiv preprint arXiv: 2002.07338*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A Other Masking Approaches Implementation Details

This section describes the implementation details of each masking approaches namely random, BM25, NOUN, TF-IDF maskings.

Random - Here we randomly mask words by adopting whole-word masking BERT (WW-BERT). Specifically, whole-word masking has known for being the standard approach to force the language model to encompass more contextual semantic dependencies.

BM25 - Here we follow the following equations to select mask words.

$$BM25(t,d) = IDF(t,D) \cdot \frac{(k+1) \cdot TF(t,d)}{k \cdot [(1-b) + b \cdot \frac{|d|}{\frac{1}{|D|}\sum_i^{|D|} |d_i|}] + TF(t,d)} \tag{4}$$

Note that b and k are parameters which are kept at 0.75 and 1.1 respectively. Below describes how $TF(t,d)$ and $IDF(t,D)$ are obtained.

$$TF(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \tag{5}$$

where $n_{t,d}$ denote number of occurrence term t in document d and $\sum_k n_{k,d}$ denote total number of keywords and documents.

$$IDF(t,D) = \log \frac{|D|}{1 + |\{d \in D : t \in d\}|} \tag{6}$$

where D denote a set of documents, d denote the current document and t denote current term.

NOUN - Specifically, we employed SpaCy (Honnibal and Montani, 2017) to identify noun in the text. Here we parsed our text to SpaCy to obtain Part-of-speech Tagging (POS) and words that are defined as Noun are selected.

TF-IDF - Here we follow the following equations to select mask words.

$$TF-IDF(t,d,D) = TF(t,d) \times IDF(t,D) \tag{7}$$

where $TF(t,d)$ and $IDF(t,D)$ are obtained the same way as those of BM25. Note that The TF-IDF value increases when a specific keyword has high frequency in a document and the frequency of documents that contain the keyword among the whole documents is low. Hence, these terms are considered relevant. Here, in this work, we refer word as term in TF-IDF.

B Masking percentage justification

	15%	30%	45%
Random	31.31	30.75	30.75
BM25	30.52	30.12	30.12
Noun	31.93	31.75	31.55
TF-IDF	31.61	30.93	30.54
PICO	32.89	32.03	31.85
Average	31.65	31.12	30.96

Table 10: Summary generation performance f1 ROUGE-1

	15%	30%	45%
Random	9.42	9.42	9.42
BM25	9.22	9.22	9.22
Noun	10.35	10.13	10.10
TF-IDF	9.68	9.51	9.31
PICO	10.79	10.46	10.25
Average	9.89	9.89	9.66

Table 11: Summary generation performance f1 ROUGE-2

	15%	30%	45%
Random	20.85	20.51	20.31
BM25	20.50	20.40	20.15
Noun	21.75	20.95	20.66
TF-IDF	20.96	20.81	20.51
PICO	21.85	21.22	21.01
Average	21.18	21.18	20.53

Table 12: Summary generation performance f1 ROUGE-L