



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

## ARCHIVIO ISTITUZIONALE DELLA RICERCA

### Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

Retrieve-and-Rank End-to-End Summarization of Biomedical Studies

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Gianluca Moro, L.R. (2023). Retrieve-and-Rank End-to-End Summarization of Biomedical Studies [10.1007/978-3-031-46994-7\_6].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/962117> since: 2024-02-26

*Published:*

DOI: [http://doi.org/10.1007/978-3-031-46994-7\\_6](http://doi.org/10.1007/978-3-031-46994-7_6)

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

# Retrieve-and-Marginalize End-to-End Summarization of Biomedical Studies

Gianluca Moro<sup>1</sup><sup>a</sup>, Luca Ragazzi<sup>1</sup><sup>b</sup>, Lorenzo Valgimigli<sup>1</sup><sup>c</sup> and Lorenzo Molfetta<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering (DISI), University of Bologna, Via dell'Università 50, Cesena, Italy  
{gianluca.moro, l.ragazzi, lorenzo.valgimigli}@unibo.it

Keywords: Multi-Document Summarization, End-to-End Learning, Biomedical Domain, Covid-19

Abstract: Real-world biomedical applications include summarizing evidence from multiple related studies given an input context to generate reviews or answer questions, which we dub context-aware multi-document summarization (CA-MDS). State-of-the-art (SOTA) solutions require truncating the input because of high memory requirements, losing potentially meaningful contents. To this end, we propose RAMSES, a retrieve-and-marginalize end-to-end summarization to effectively address biomedical CA-MDS tasks by marginalizing at decoding time the token probability distributions of latent retrieved documents conditioned by an input context. For evaluation, we present a new dataset, FAQSUMC19, to answer questions on Covid-19 by synthesizing multiple supporting papers. Results reveal that RAMSES achieves significantly higher ROUGE scores than cutting-edge methods, including a new SOTA on MS2 for generating systematic literature reviews. Human evaluation also reports that our model generates more informative answers than the previous SOTA approach.

## 1 INTRODUCTION

Given the paramount societal role of biomedicine, natural language processing (NLP) tasks for aggregating information from multiple topic-related biomedical papers to help search, synthesize, and answer questions are of high interest (DeYoung et al., 2021). Real-world applications require combining and summarizing evidence from clinical trials on a research background to produce systematic literature reviews (SLRs) or answer medical inquiries. Consequently, we define such activities as *context-aware multi-document summarization* (CA-MDS) because of the presence of an input context (i.e., the background or question) that conditions the downstream summarization task (Figure 1). In real life, biomedical articles usually contain several thousands of words penning lingo and complicated expressions, making understanding them a time- and labor-demanding process even for professionals. For this reason, automation support for complex biomedical activities is practical and beneficial to facilitate knowledge acquisition.

CA-MDS solutions for biomedical applications should process all inputs without ignoring any de-

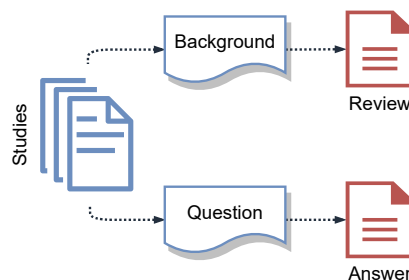





Figure 1: The overview of biomedical CA-MDS. In our experiments, we use the input contexts (i.e., the background or question) to retrieve the salient studies and aggregate them to generate the target (Input → Output).

tails, reducing the risk of model hallucination, namely generating unfaithful outputs due to training on targets having facts unfounded by the source. So, state-of-the-art (SOTA) models rely on sparse transformers (Beltagy et al., 2020), Fusion-in-Decoder strategies (Izacard and Grave, 2021a), and marginalization-based decoding (Moro et al., 2022a). Yet, such methods either (i) need high memory requirements forcing input truncation for those organizations that operate in low-resource regimes (Moro and Ragazzi, 2022; Moro et al., 2023b; Moro and Ragazzi, 2023), or (ii) lack end-to-end learning diminishing the potential of cooperating neural modules.

In this paper, we introduce RAMSES, a retrieve-

<sup>a</sup> <https://orcid.org/0000-0002-3663-7877>

<sup>b</sup> <https://orcid.org/0000-0003-3574-9962>

<sup>c</sup> <https://orcid.org/0000-0003-0309-771X>

and-marginalize summarization approach trained via end-to-end learning to synthesize the knowledge from numerous topic-related biomedical documents given an input context. RAMSES comprises a biomedical bi-encoder and a generative aggregator. The bi-encoder retrieves and scores salient documents related to an input context. Then, the aggregator is conditioned by the context along with these latent documents to decode the summary by marginalizing the token probability distribution weighted by their relevance score.

We assess RAMSES in two biomedical CA-MDS tasks: (i) producing SLRs on the Ms2 dataset (DeYoung et al., 2021) and (ii) answering frequently asked questions (FAQs) about Covid-19 on our proposed dataset FAQSUMC19.<sup>1</sup> In detail, we collected 514 Covid-19 FAQs with high-quality expert-written abstractive answers. Then, we augmented each instance with 30 supporting scientific papers containing needed information to answer the question, yielding 15,420 articles. Notably, FAQSUMC19 has two essential features: (i) it includes expert-authored abstractive answers unlike other related datasets that use extractive targets (Rajpurkar et al., 2018); (ii) it is the first CA-MDS dataset for Covid-19, becoming a crucial benchmark for producing multi-document summaries to answer questions on Covid-19 with the support of updated related biomedical papers.

We accomplish extensive experiments, showing that RAMSES achieves new SOTA performance on the Ms2 dataset and outperforms previous solutions on FAQSUMC19, whose inferred answers are also rated as of more quality by human experts.

## 2 RELATED WORK

**NLP for Biomedical Documents.** Much recent work in NLP has concentrated on the biomedical domain (Domeniconi et al., 2016; Moro and Masseroli, 2021), including CA-MDS (DeYoung et al., 2021), which can decrease the burdens on medical workers by highlighting and aggregating key points while reducing the amount of information to read. Previous contributions focused on the automatic generation of SLRs. In detail, cutting-edge solutions rely on three different neural architectures: (i) transformer-based models with linear complexity in the input size thanks to sparse attention (Beltagy et al., 2020), which concatenate the input context along with all documents in the cluster producing a single source sequence; (ii) quadratic transformers with Fusion-in-Decoder (Izacard and Grave, 2021b), which join the hidden states

of documents after encoding them individually; (iii) marginalization-based decoding augmented by frozen retrievers (Moro et al., 2022a), which first pinpoints salient documents w.r.t. a query and separately produces a single summary by summing the probability distribution of the inferred token for each document.

**MDS Solutions in other Domains.** Flat approaches with MDS-specific pre-training (Xiao et al., 2022) concatenate the sources in a single text, treating MDS as a single-input task. Hierarchical approaches merge document relations to obtain semantic-rich representations by leveraging graph-based methods (Amplayo and Lapata, 2021) and multi-head pooling and inter-paragraph attention (Jin et al., 2020). Marginalization-based approaches (Hokamp et al., 2020) apply marginalization to the token probability distribution at decoding time to produce a single output from many inputs. Two-stage approaches (Liu and Lapata, 2019) adopt different strategies to rank sources before producing the summary. Unlike prior works, RAMSES is trained in end-to-end learning to retrieve salient text across biomedical articles and marginalize the probability distribution of the latent extracted information at decoding time.

**Covid-19 Datasets.** With the Covid-19 appearance, thousands of papers have been released quickly. To aid experts access this knowledge, large organizations collected corpora like CORD-19 (Wang et al., 2020) and LITCOVID (Chen et al., 2021), fostering the proposal of task-specific datasets. COVID-QA (Möller et al., 2020) study question-answering using annotated pairs extracted from 147 papers. COVID-Q (Wei et al., 2020) collects 16,690 questions about Covid-19, classifying them into 15 categories. (Poliak et al., 2020) scrapped over 40 trusted websites of FAQs about Covid-19, creating a collection of 2100 questions. (Zhang et al., 2021) and (Sun and Sedoc, 2020) proposed two datasets for FAQ retrieval, where user queries are semantically paired with existing FAQs. FAQSUMC19 fills this gap, introducing the first CA-MDS dataset for answering Covid-19 FAQs by summarizing multiple related studies.

## 3 PRELIMINARY

We provide details about the task of context-aware multi-document summarization (CA-MDS).

**Definition.** CA-MDS aims to assemble a summary from a cluster of related articles given an input context—analogue to the query in query-focused summarization (Vig et al., 2022). Yet, unlike FAQ answering, the SLR generation does not consider questions. Thus, we define the task we face as CA-MDS.

<sup>1</sup><https://github.com/disi-unibo-nlp/faqsumc19>

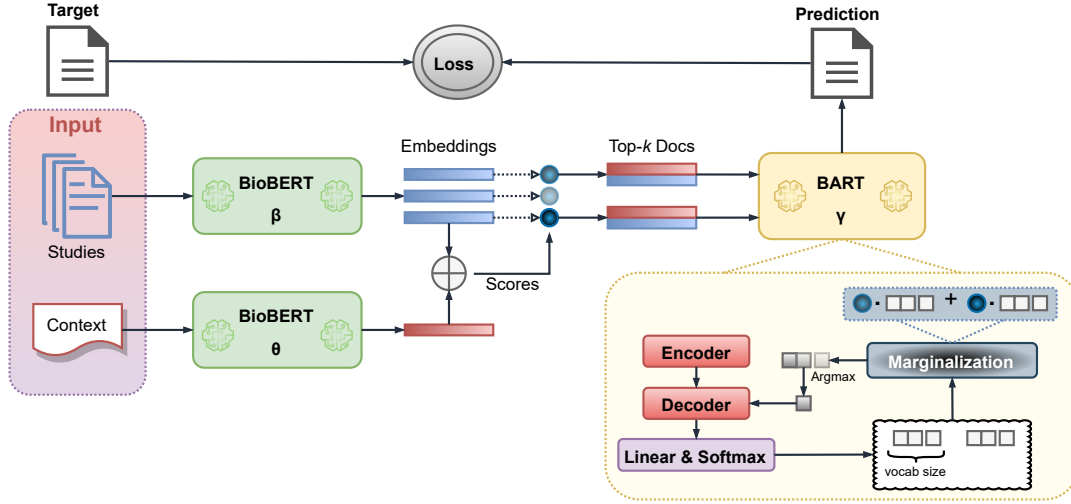


Figure 2: The overview of RAMSES. The input comprises a biomedical context and multiple related studies that are encoded by two different BIOBERT models. Then, we compute the relevance score of each document conditioned by the context. Finally, the top- $k$  most salient documents are concatenated with the context and given to a BART model that marginalizes their token probability distribution, weighted by the relevance scores, at decoding time.

The biomedical tasks we address in this work—such as SLR generation and FAQ answering—are CA-MDS tasks because they both have an input context (i.e., the research issue in SLRs and the human question in FAQs) and many topic-related documents from which produce the output.

**Problem Formulation.** In the CA-MDS setting, we have  $(c, \mathbf{D}, y)$ , where  $c$  is the input context,  $\mathbf{D}$  is the cluster of topic-related documents, and  $y$  is the target generated from  $\mathbf{D}$  given  $c$ . Formally, we want to predict  $y$  from  $\{c, d_1, \dots, d_n | d \in \mathbf{D}\}$ .

## 4 METHOD

The end-to-end learning of RAMSES allows its cooperating modules to jointly retrieve and aggregate key information across sources in one output (Figure 2).

Given the context  $c$  and documents  $\mathbf{D}$ , our method first generates relevance scores over  $\mathbf{D}$  with a biomedical DPR-based solution (Karpukhin et al., 2020):

$$p_{\beta, \theta}(d \in \mathbf{D} | c) = (Enc_{\beta}(d) \oplus Enc_{\theta}(c)) \quad (1)$$

where  $Enc_{\beta}$  and  $Enc_{\theta}$  are two different BIOBERT-base models trained to produce a dense representation of documents and the context (Papanikolaou and Bennett, 2021), respectively,  $\oplus$  is the inner product between them, and  $p(d|c)$  is the relevance score associated to the document  $d$  given  $c$ . Thus, our solution finds the most top- $k$  relevant texts according to  $c$ . Then, given  $c$  and each  $d \in \text{top-}k$ , a BART-base model

(Lewis et al., 2020) draws a distribution for each next output token for each  $d$ , before marginalizing:

$$p(y|c, \mathbf{D}) = \prod_z \sum_{d \in \text{top-}k} p_{\theta}(d|c) p_{\gamma}(y_z | d', y_{1:z-1}) \quad (2)$$

where  $d' = [c, tok, d]$  is the concatenation of  $c$  and  $d \in \text{top-}k$  with a special text separator token ( $\langle \text{doc-sep} \rangle$ ) to make the model aware of the textual boundary,  $N$  is the target length, and  $p_{\gamma}(y_z | d', y_{1:z-1})$  is the probability of generating the target token  $y_z$  given  $d'$  and the previously generated tokens  $y_{1:z-1}$ .

We train our RAMSES model by minimizing the negative marginal log-likelihood of each target with the following loss function:

$$\mathcal{L} = -\sum_i \log p(y_i | c_i, \mathbf{D}_i) \quad (3)$$

### 4.1 End-to-End Learning

The model (Equation 2) allows the gradient to back-propagate to all modules. For clarity, we rewrite the formula as a continuous function, as follows:

$$\text{RAMSES}(\mathbf{D}, c) = \sum_{(d_j, s_j)} B_{\gamma}([c, tok, d_j]) \cdot s_j \quad (4)$$

$$\text{top-}k(\mathbf{D}, c) = [(d_1, s_1), \dots, (d_k, s_k)] \quad (5)$$

$$s_j = Enc_{\beta}(d_j) \oplus Enc_{\theta}(c) \quad (6)$$

where  $(d_j, s_j) \in \text{top-}k$  and  $B_{\gamma}$  is BART.

The presence of  $s_j$  in Equation 4 lets the gradient, computed by minimizing the objective function, reach

Table 1: The question-cluster pairs’ quality with ROUGE-1 precision and BERTScore. Best values are bolded.

	ROUGE			BERTScore		
	Avg	Max	Min	Avg	Max	Min
RANDOM	12.34	23.21	0.17	11.30	24.71	2.80
BM25	16.70	29.68	0.37	16.17	35.20	1.13
SUBLIMER	<b>20.44</b>	<b>33.31</b>	<b>2.32</b>	<b>21.11</b>	<b>36.79</b>	<b>5.75</b>

$Enc_{\beta}$  and  $Enc_{\theta}$ . For this reason, the documents and context embeddings are adjusted during the train to improve the generated summary, making all modules of our solution learn jointly in an end-to-end fashion.

## 5 FAQSUMC19 DATASET

We introduce a new dataset, FAQSUMC19, containing 514 FAQs related to Covid-19 with expert-written abstractive answers, each supported by 30 scientific paper abstracts, for a total of 15,420 articles. We fetched from the Covid-19 FAQ section on WHO<sup>2</sup> all question-answer pairs available. We then augmented each instance with 30 Covid-19 scientific articles strictly related to the question from the updated version of the CORD-19 dataset (Wang et al., 2020). Precisely, we experimented the selection of the supporting papers with different information retrieval methods, such as a random baseline, BM25 (Robertson and Walker, 1994), and SUBLIMER (Moro and Valgimigli, 2021). We used the concatenation between the question and the answer to retrieve the first 30 ranked documents in terms of semantic similarity, creating a knowledge basis to support the answer generation. We finally split the dataset into 464 instances for training ( $\approx 90\%$ ) and 50 for test ( $\approx 10\%$ ).

To assess the question-cluster pairs’ quality of our dataset, we computed the content coverage with ROUGE-1 precision (Lin, 2004) and BERTScore (Zhang et al., 2020b) of the question-answer (q-a) concatenation w.r.t. each document in the cluster, and calculate the average score. So, we evaluate the syn-

Table 2: The datasets used for evaluation (FAQSUMC19 is our proposed dataset). Statistics include dataset size and the average (i) number of source (S) documents per instance, (ii) number of total words in S and target (T) texts, and (iii) S-T compression ratio of words (Grusky et al., 2018).

Dataset	Samples	S		T	S $\rightarrow$ T
		Docs	Words	Words	Comp
Ms2	15,597	23.30	9563.95	70.81	135.06
FAQSUMC19	514	30	5635.50	139.06	40.53

<sup>2</sup><https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub>

tactic and semantic overlap between q-a and the supporting texts. Table 1 reports the results and reveals that SUBLIMER achieves the best scores, as expected.

## 6 EXPERIMENTS

### 6.1 Experiment Setup

**Datasets.** Table 2 reports the statistics of datasets used to test RAMSES on different biomedical tasks: **Ms2** (DeYoung et al., 2021) consists of 15,597 instances derived from the scientific literature. Each sample is composed of: (i) the background statement, which describes the context research issue, (ii) the target statement, which is the summary to generate, and (iii) the studies, which are the abstracts of biomedical documents that contain the needed information for the research issue. **FAQSUMC19** is our proposed dataset that comprises 514 FAQs related to Covid-19 with expert-written abstractive answers, each supported by 30 scientific paper abstracts.

**Baselines.** We compare RAMSES with SOTA solutions: **BART-FID** (DeYoung et al., 2021), which is BART with the Fusion-in-Decoder strategy (Izard and Grave, 2021b), encodes all sources individually and combines their hidden states before decoding. **LED-GAQ** (DeYoung et al., 2021), which is LED (Beltagy et al., 2020) with global attention on the input query, concatenates all texts in a single input up to 16,384 tokens. **DAMEN** (Moro et al., 2022a), a retrieval-enhanced solution with a marginalization-based decoding, discriminates important snippets from the cluster with a frozen BERT-based model and marginalizes their probability distribution during decoding. **PRIMERA** (Xiao et al., 2022), which is LED pre-trained with a multi-document summarization-specific objective, concatenates the texts with a special separator token up to 4096 tokens in size.

**Evaluation Metrics.** We use ROUGE-1/2/L (Lin, 2004) to assess fluency and informativeness. We also adopt  $\mathcal{R}$  (Moro et al., 2023a) as an aggregated judgment that considers the variance of ROUGE scores. Finally, we accomplish qualitative analysis to fill the superficiality of automatic evaluation measures.

**Implementation.** We fine-tune the models using PyTorch and the HuggingFace library, setting the seed to 42 for reproducibility. RAMSES is trained on one NVIDIA RTX 3090 GPU of 24 GB memory from an internal cluster for 1 epoch with a learning rate of  $3e-5$  on Ms2 and for 3 epochs with a learning rate of  $1e-5$  on FAQSUMC19. For decoding, we use beam search with 4 beams and the following min-max target size: 32-256 for Ms2 and 100-256 for FAQSUMC19.

Table 3: Performance of models on Ms2 and FAQSUMC19 evaluation datasets. Best scores are in bold.

Model	Ms2				FAQSUMC19			
	R-1 <sub>f1</sub>	R-2 <sub>f1</sub>	R-L <sub>f1</sub>	$\mathcal{R}$	R-1 <sub>f1</sub>	R-2 <sub>f1</sub>	R-L <sub>f1</sub>	$\mathcal{R}$
<b>Baselines</b>								
LED-GAQ	26.89	8.91	20.32	18.60	25.55	4.42	13.77	14.47
BART-FID	27.56	9.49	20.80	19.18	20.26	5.59	14.84	13.51
DAMEN	28.95	9.72	21.83	20.04	23.81	3.50	13.03	13.35
PRIMERA	30.07	9.85	22.16	20.55	25.04	3.64	13.00	13.79
<b>Our</b>								
RAMSES	<b>31.83</b>	<b>10.44</b>	<b>22.19</b>	<b>21.32</b>	<b>30.18</b>	<b>7.31</b>	<b>15.67</b>	<b>17.56</b>

Table 4: ROUGE F1 scores (R-1, R-2, R-L) on Ms2 on evaluating RAMSES with different generator checkpoints (B and L stand for base and large, respectively) and  $k$  documents retrieved at training time. OOM means ‘‘GPU out of memory exception’’. Best results are bolded.

$k$	Ms2		
	BART-B	BART-L	PEGASUS-L
3	31.14/9.78/21.43	31.93/10.40/21.96	27.83/7.27/18.93
6	31.12/9.88/21.68	<b>31.97/10.58/22.15</b>	<u>28.71/8.26/19.39</u>
9	<u>31.81/10.30/22.13</u>	31.00/10.17/21.76	28.61/8.32/19.35
12	31.15/10.14/21.58	31.10/10.09/6.52	27.58/7.30/18.51
15	30.94/9.82/21.39	31.72/10.53/21.90	OOM
18	31.36/10.20/21.75	31.81/10.43/21.87	OOM

## 6.2 Results

Table 3 reports the performance of models on the two evaluation datasets. RAMSES yields better scores, suggesting that the retrieve-and-marginalize end-to-end learning is more effective than prior SOTA approaches on both biomedical CA-MDS tasks.

### 6.2.1 The Impact of $k$

As our method relies on learning to select the best top- $k$  relevant documents from the cluster, the value of  $k$  is crucial for model performance and GPU memory occupation. Therefore, we analyze the impact of  $k$  on model performance by experimenting with a different number of documents to retrieve: 3, 6, 9, 12, 15, 18. Table 4 reports a slight performance improvement as  $k$  increases until a threshold is reached (e.g.,  $k = 9$  for BART-base), indicating that the marginalization approach with more documents helps to produce better ROUGE scores. However, a high  $k$  (i.e.,  $k \geq 12$ ) can also increase information redundancy and contradiction, lowering the final performance.

Table 4 also lists the results of different models on single text summarization as the aggregator’s checkpoint, such as BART and PEGASUS (Zhang et al., 2020a). We notice that BART-large achieves better ROUGE scores, although PEGASUS is the largest model. Yet, as BART-base achieved a slightly lower result despite the noticeably fewer trainable param-

Table 5: The results of RAMSES on Ms2 by varying  $k$  at inference time and on FAQSUMC19 by varying  $k$  at training time. Best scores are bolded.

$k$	Ms2			FAQSUMC19		
	R-1 <sub>f1</sub>	R-2 <sub>f1</sub>	R-L <sub>f1</sub>	R-1 <sub>f1</sub>	R-2 <sub>f1</sub>	R-L <sub>f1</sub>
3	30.98	9.75	21.48	29.32	7.05	15.71
6	31.49	10.14	21.94	28.69	6.51	15.20
9	31.81	10.30	22.13	28.74	6.83	15.44
12	<b>31.83</b>	<b>10.44</b>	<b>22.19</b>	<b>30.18</b>	<b>7.31</b>	<b>15.67</b>
15	31.73	10.36	22.10	29.14	6.65	15.26
18	31.72	10.39	22.04	29.06	6.89	15.31

ters, we chose to use it for all experiments. Therefore, we tested the best checkpoint of BART-base trained with  $k = 9$  with a different  $k$  at inference time on Ms2. Table 5 reports that the best performance has been achieved with  $k = 12$ . Furthermore, Table 5 also shows the results on FAQSUMC19 with a different  $k$  at training time, revealing a similar trend to Ms2.

**Memory Requirements.** Figure 3 shows the memory complexity at training time of RAMSES for each  $k$ . We notice that the memory occupation is linear w.r.t.  $k$ , indicating that our solution is not computationally expensive, even for large clusters.

## 6.3 Ablation Studies

Table 6 reports the ablation studies on Ms2 using RAMSES with BART-base and  $k = 9$  with the same hyperparameter settings for all experiments.

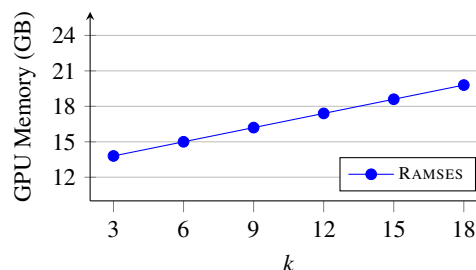


Figure 3: The RAMSES’s GPU memory requirements by varying  $k$  at training time.

Table 6: The ablation studies on Ms2. We gradually remove each module of RAMSES to show the performance drop. Best scores are in bold.

	Ms2			
	R-1 <sub>f1</sub>	R-2 <sub>f1</sub>	R-L <sub>f1</sub>	$\mathcal{R}$
RAMSES	<b>31.83</b>	<b>10.44</b>	<b>22.19</b>	<b>21.32</b>
w/o context-first	31.57	10.26	21.89	21.08
w/o trained-retrieval	31.24	10.03	21.77	20.86
w/o inner-product	31.12	10.10	21.82	20.86
w/o token-sep	31.03	10.12	21.77	20.82
w/o bi-encoder	31.47	9.88	21.52	20.79
w/o context	25.26	5.33	17.37	15.88

Excluding the input context from the input concatenation to give to the generative aggregator (**w/o context**) leads to the most significant decrease in performance. Indeed, the context is the research question shared by all documents in the cluster, so it contains important information for the final summary.

Training a single model to encode both the context and documents (**w/o bi-encoder**), i.e., using a shared BIOBERT model, decreases performance. Indeed, since the context and the documents have two different purposes (i.e., we need the context to select context-related documents), two models are needed to specialize and differentiate the text representation. Furthermore, despite the similarity of the two tasks, they are different for two main reasons: (i) the context is relatively shorter than the documents in the cluster and (ii) the concept expressiveness is denser in the context than in the more verbose documents.

Removing the token separator `<doc-sep>` between the context and document (**w/o token-sep**) decreases performance. Indeed, this token is needed to make the model aware of the textual boundary between the context and the documents.

Using cosine similarity instead of the inner product (**w/o inner-product**) to score the documents against the input context achieves worst results.

Freezing  $Enc_{\beta}$  (**w/o trained-retrieval**) decreases performance, highlighting the helpfulness of end-to-end learning to let the model select more informative documents in the cluster.

Switching the context with the documents in the input concatenation (**w/o context-first**) decreases performance, indicating that the context at the beginning of the input helps the generative aggregator to focus better on how to join context-related information.

## 6.4 Human Evaluation

Considering the drawbacks of automatic metrics such as ROUGE (Fabbri et al., 2021), which is still the standard for evaluating text generation, we qualitatively assessed the inferred FAQs answers of the en-

Table 7: The human evaluation on FAQSUMC19 with inter-annotator agreement (IAA) using WHO ground-truth answers and the inferred ones by RAMSES and LED-large. A>B means how many times the generated answer from A was scored higher than B.

	Human Evaluation		
	RAMSES>LED	RAMSES>WHO	LED>WHO
Evaluator 1	84%	0%	0%
Evaluator 2	72%	20%	14%
Evaluator 3	72%	2%	0%
AVG	76%	7.33%	4.67%
IAA	46%	80%	86%

tire FAQSUMC19 test set with three domain experts with master degrees in medical and biological areas.

**Instructions.** We gave evaluators a table, with each row containing the question and three possible answers in random order: (i) the “gold” from WHO, (ii) the prediction of RAMSES, and (iii) the prediction of LED-large (the second-best model on FAQSUMC19 according to  $\mathcal{R}$ ). Each expert was asked to order the answers according to how thoroughly they answered the question, primarily focusing on factuality. For fairness, we did not inform evaluators about the answers’ origins and the test’s goal. Overall, experts completed the task in two days, reporting no difficulty in ordering the 50 answers.

**Results.** Evaluation results, reported in Table 7, show that our method produces better informative abstractive answers to a given open question than a linear transformer with sparse attention. Precisely, experts rated 76% of the answers from our solution to be better than LED’s ones, with an inter-annotator agreement of 46% (meaning that 46% of the time all three evaluators agree). Moreover, evaluators also found that 7.33% of the answers inferred by our RAMSES model are more informative than the “gold” from WHO. Nevertheless, the model generations are still far from being informative as the “gold” answers, indicating the limitations of current neural language models on FAQSUMC19.

## 7 CONCLUSION

In this paper, we introduced RAMSES, a retrieve-and-marginalize end-to-end learning solution for CAMDS of biomedical studies. Through multiple experiments on two biomedical datasets (including our proposed FAQSUMC19 for Covid-19 FAQ answering), we found that RAMSES outperforms SOTA models, even though there is still significant room for improvement, as suggested by human assessments. We hope this work can foster research toward new reli-

able solutions for biomedical applications.

For future works, we indicate investigating: (i) memory-based operations from unsupervised approaches for classes extraction (Domeniconi et al., 2014; Domeniconi et al., 2015; Domeniconi et al., 2017; Moro et al., 2018) and entity relationships acquisition (Frisoni et al., 2020; Frisoni and Moro, 2020) to avant-garde semantic parsing solutions such as event extraction (Frisoni et al., 2021; Frisoni et al., 2022c) and abstract meaning representation (Frisoni et al., 2022a; Frisoni et al., 2023); (ii) retrieval-enhanced techniques (Moro et al., 2022b; Frisoni et al., 2022b). New directions should consider involving novel graph representation learning methods (Ferrari et al., 2022; Frisoni et al., 2022d). Lastly, as proposed for communication networks (Monti and Moro, 2008; Lodi et al., 2010; Moro and Monti, 2012; Cerroni et al., 2013; Cerroni et al., 2015), tracking and propagating knowledge refinements across sentences could be critical when tackling extended sequences.

## REFERENCES

- Amplayo, R. K. and Lapata, M. (2021). Informative and controllable opinion summarization. In *EACL, Online, April 19-23 2021*, pages 2662–2672. ACL.
- Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *CoRR*, abs/2004.05150.
- Cerroni, W., Moro, G., Pasolini, R., and Ramilli, M. (2015). Decentralized detection of network attacks through P2P data clustering of SNMP data. *Comput. Secur.*, 52:1–16.
- Cerroni, W., Moro, G., Pirini, T., and Ramilli, M. (2013). Peer-to-peer data mining classifiers for decentralized detection of network attacks. In *ADC*, volume 137 of *CRPIT*, pages 101–108. ACS.
- Chen, Q., Allot, A., and Lu, Z. (2021). Litcovid: an open database of COVID-19 literature. *Nucleic Acids Res.*, 49(Database-Issue):D1534–D1540.
- DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., et al. (2021). Ms<sup>2</sup>: Multi-document summarization of medical studies. In *EMNLP, Punta Cana, 7-11 November, 2021*, pages 7494–7513. ACL.
- Domeniconi, G., Masseroli, M., Moro, G., and Pinoli, P. (2016). Cross-organism learning method to discover new gene functionalities. *Comput. Methods Programs Biomed.*, 126:20–34.
- Domeniconi, G., Moro, G., Pagliarani, A., and Pasolini, R. (2015). Markov chain based method for in-domain and cross-domain sentiment classification. In *KDIR*, pages 127–137. SciTePress.
- Domeniconi, G., Moro, G., Pagliarani, A., and Pasolini, R. (2017). On deep learning in cross-domain sentiment classification. In *IC3K, Funchal, Madeira, Portugal, November 1-3, 2017*, pages 50–60. SciTePress.
- Domeniconi, G., Moro, G., Pasolini, R., and Sartori, C. (2014). Cross-domain text classification through iterative refining of target categories representations. In *KDIR, Rome, Italy, 21 - 24 October, 2014*, pages 31–42. SciTePress.
- Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., et al. (2021). Summeval: Re-evaluating summarization evaluation. *TACL*, 9:391–409.
- Ferrari, I., Frisoni, G., Italiani, P., Moro, G., and Sartori, C. (2022). Comprehensive analysis of knowledge graph embedding techniques benchmarked on link prediction. *Electronics*, 11(23).
- Frisoni, G., Italiani, P., Boschi, F., and Moro, G. (2022a). Enhancing biomedical scientific reviews summarization with graph-based factual evidence extracted from papers. In *DATA, Lisbon, Portugal, July 11-13, 2022*, pages 168–179. SCITEPRESS.
- Frisoni, G., Italiani, P., Salvatori, S., and Moro, G. (2023). Cogito ergo summ: Abstractive summarization of biomedical papers via semantic parsing graphs and consistency rewards. In *AAAI*, pages 1–9. AAAI Press.
- Frisoni, G., Mizutani, M., Moro, G., and Valgimigli, L. (2022b). Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *EMNLP 2022*, pages 5770–5793, Abu Dhabi, United Arab Emirates. ACL.
- Frisoni, G. and Moro, G. (2020). Phenomena Explanation from Text: Unsupervised Learning of Interpretable and Statistically Significant Knowledge. In *DATA*, volume 1446, pages 293–318. Springer.
- Frisoni, G., Moro, G., and Balzani, L. (2022c). Text-to-text extraction and verbalization of biomedical event graphs. In *COLING*, pages 2692–2710, Gyeongju, Republic of Korea. ICCL.
- Frisoni, G., Moro, G., and Carbonaro, A. (2020). Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining. In *DATA*, pages 121–134. SciTePress.
- Frisoni, G., Moro, G., and Carbonaro, A. (2021). A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757.
- Frisoni, G., Moro, G., Carlassare, G., and Carbonaro, A. (2022d). Unsupervised event graph representation and similarity learning on biomedical literature. *Sensors*, 22(1):3.
- Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *NAACL (Long Papers)*, pages 708–719, New Orleans, Louisiana. ACL.
- Hokamp, C., Ghalandari, D. G., Pham, N. T., and Glover, J. (2020). Dyne: Dynamic ensemble decoding for multi-document summarization. *CoRR*, abs/2006.08748.
- Izcard, G. and Grave, E. (2021a). Leveraging passage retrieval with generative models for open domain question answering. In *EACL: Main Volume*, pages 874–880, Online. ACL.



- Izacard, G. and Grave, E. (2021b). Leveraging passage retrieval with generative models for open domain question answering. In *EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. ACL.
- Jin, H., Wang, T., and Wan, X. (2020). Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *ACL, Online, July 5-10 2020*, pages 6244–6254. ACL.
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., et al. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., et al. (2020). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL, July 5-10 2020*, pages 7871–7880.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.
- Liu, Y. and Lapata, M. (2019). Hierarchical transformers for multi-document summarization. In *ACL, Florence, Italy, July 28- August 2 2019*, pages 5070–5081. ACL.
- Lodi, S., Moro, G., and Sartori, C. (2010). Distributed data clustering in multi-dimensional peer-to-peer networks. In *(ADC), Brisbane, 18-22 January, 2010*, volume 104 of *CRPIT*, pages 171–178. ACS.
- Möller, T., Reina, A., Jayakumar, R., and Pietsch, M. (2020). Covid-qa: a question answering dataset for covid-19.
- Monti, G. and Moro, G. (2008). Multidimensional range query and load balancing in wireless ad hoc and sensor networks. In *Peer-to-Peer Computing*, pages 205–214. IEEE Computer Society.
- Moro, G. and Masseroli, M. (2021). Gene function finding through cross-organism ensemble learning. *BioData Min.*, 14(1):14.
- Moro, G. and Monti, G. (2012). W-grid: A scalable and efficient self-organizing infrastructure for multi-dimensional data management, querying and routing in wireless data-centric sensor networks. *J. Netw. Comput. Appl.*, 35(4):1218–1234.
- Moro, G., Pagliarani, A., Pasolini, R., and Sartori, C. (2018). Cross-domain & In-domain Sentiment Analysis with Memory-based Deep Neural Networks. In *IC3K 2018*, volume 1, pages 127–138. SciTePress.
- Moro, G. and Ragazzi, L. (2022). Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *AAAI 2022, Virtual Event, February 22 - March 1, 2022*, pages 11085–11093. AAAI Press.
- Moro, G. and Ragazzi, L. (2023). Align-then-abstract representation learning for low-resource summarization. *Neurocomputing*, 548:126356.
- Moro, G., Ragazzi, L., and Valgimigli, L. (2023a). Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. *AAAI*, 37(12):14417–14425.
- Moro, G., Ragazzi, L., Valgimigli, L., and Freddi, D. (2022a). Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In *ACL*, pages 180–189, Dublin, Ireland. ACL.
- Moro, G., Ragazzi, L., Valgimigli, L., Frisoni, G., Sartori, C., and Marfia, G. (2023b). Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. *Sensors*, 23(7).
- Moro, G. and Valgimigli, L. (2021). Efficient self-supervised metric information retrieval: A bibliography based method applied to COVID literature. *Sensors*, 21(19).
- Moro, G., Valgimigli, L., Rossi, A., Casadei, C., and Montefiori, A. (2022b). Self-supervised information retrieval trained from self-generated sets of queries and relevant documents. In *SISAP, Bologna, Italy, October 5-7, 2022, Proceedings*, volume 13590 of *Lecture Notes in Computer Science*, pages 283–290. Springer.
- Papanikolaou, Y. and Bennett, F. (2021). Slot filling for biomedical information extraction. *CoRR*, abs/2109.08564.
- Poliak, A., Fleming, M., Costello, C., Murray, K. W., et al. (2020). Collecting verified COVID-19 question answer pairs. In *NLP4COVID@EMNLP*. ACL.
- Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don’t know: Unanswerable questions for squad. In *ACL 2018, Melbourne, Australia, July 15-20, 2018*, pages 784–789. ACL.
- Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *ACM-SIGIR. Dublin, 3-6 July 1994*, pages 232–241. ACM/Springer.
- Sun, S. and Sedoc, J. (2020). An analysis of bert faq retrieval models for covid-19 infobot.
- Vig, J., Fabbri, A. R., Kryscinski, W., Wu, C., et al. (2022). Exploring neural models for query-focused summarization. In *NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1455–1468. ACL.
- Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., et al. (2020). COR-19: the covid-19 open research dataset. *CoRR*, abs/2004.10706.
- Wei, J. W., Huang, C., Vosoughi, S., and Wei, J. (2020). What are people asking about covid-19? A question classification dataset. *CoRR*, abs/2005.12522.
- Xiao, W., Beltagy, I., Carenini, G., and Cohan, A. (2022). PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *ACL*, pages 5245–5263, Dublin. ACL.
- Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020a). PE-GASUS: pre-training with extracted gap-sentences for abstractive summarization. In *ICML, 13-18 July 2020*, volume 119, pages 11328–11339. PMLR.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., et al. (2020b). Bertscore: Evaluating text generation with BERT. In *ICLR, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Zhang, X. F., Sun, H., Yue, X., Lin, S. M., et al. (2021). COUGH: A challenge dataset and models for COVID-19 FAQ retrieval. In *EMNLP 2021, Virtual Event, 7-11 November, 2021*, pages 3759–3769. ACL.