




Discriminative Marginalized Probabilistic Neural Method for Multi-Document Summarization of Medical Literature

Gianluca Moro ^{*,†}, Luca Ragazzi ^{*,*}, Lorenzo Valgimigli ^{*,*}, Davide Freddi

^{*}Department of Computer Science and Engineering (DISI)

University of Bologna, Cesena Campus, [†]CNIT

Via dell'Università 50, I-47522 Cesena, Italy

{gianluca.moro, l.ragazzi, lorenzo.valgimigli}@unibo.it - davide.freddi3@studio.unibo.it

Abstract

Although current state-of-the-art Transformer-based solutions succeeded in a wide range for single-document NLP tasks, they still struggle to address multi-input tasks such as multi-document summarization. Many solutions truncate the inputs, thus ignoring potential summary-relevant contents, which is unacceptable in the medical domain where each information can be vital. Others leverage linear model approximations to apply multi-input concatenation, worsening the results because all information is considered, even if it is conflicting or noisy with respect to a shared background. Despite the importance and social impact of medicine, there are no ad-hoc solutions for multi-document summarization. For this reason, we propose a novel discriminative marginalized probabilistic method (DAMEN) trained to discriminate critical information from a cluster of topic-related medical documents and generate a multi-document summary via token probability marginalization. Results prove we outperform the previous state-of-the-art on a biomedical dataset for multi-document summarization of systematic literature reviews. Moreover, we perform extensive ablation studies to motivate the design choices and prove the importance of each module of our method.¹

1 Introduction

The task of multi-document summarization aims to generate a compact and informative summary from a cluster of topic-related documents, which represents a very challenging natural language processing (NLP) application due to the presence of redundant and sometimes conflicting information among documents (Radev, 2000). In the medical domain, in which machine learning plays an increasingly significant role (Domeniconi et al., 2014a; di Lena et al., 2015), multi-document summarization finds application in the generation of

systematic literature reviews, a biomedical paper that summarizes results across many studies (Khan et al., 2003). DeYoung et al. (2021) are the first that address this task, showing the related issues.

State-of-the-art approaches leverage two leading solutions: hierarchical networks that capture cross-document relations via graph encodings (Wan and Yang, 2006; Liao et al., 2018; Li et al., 2020; Pasunuru et al., 2021) or hidden states aggregation (Fabbri et al., 2019; Liu and Lapata, 2019a; Jin et al., 2020), and long-range neural models that apply multi-input concatenation (Xiao et al., 2021). While effective, these solutions struggle to process clusters of many topic-related documents in low computational resource scenarios (Moro and Ragazzi, 2022) because they need to truncate the inputs. Moreover, pre-trained state-of-the-art Transformers are not leveraged despite showing strong performance when fine-tuned in downstream tasks such as single-document summarization (Liu and Lapata, 2019b; Lewis et al., 2020a; Raffel et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020).

Multi-document summarization requires models to have more robust capabilities for analyzing the cluster to discriminate the correct information from noise and merge it consistently. In this work, we propose a discriminative marginalized probabilistic neural method (DAMEN) that selects worthy documents in the cluster with respect to a shared background and generates the summary via token probability marginalization.

The marginalization of the probability has been successfully applied in past NLP models such as pLSA (Hofmann, 1999) to learn the word probability distribution in documents by maximizing the likelihood. Recently, new deep neural models that use probability marginalization approaches have been proposed for the question-answering task, where, however, each input/output sentence is generally several orders of magnitude shorter than sets of documents in multi-document summarization

¹The solution of this paper is available at <https://disi-unibo-nlp.github.io/projects/damen>

(Guu et al., 2020; Lewis et al., 2020b).

To the best of our knowledge, we are the first that propose such a method for multi-document summarization. To this aim, we conduct experiments on the only medical dataset for multi-document summarization of systematic literature reviews (MS2). Besides, we perform extensive ablation studies to motivate the design choices and prove the importance of each component of our method.

To sum up, our contributions are as follows:

- We propose a novel probabilistic neural method for multi-document summarization (DAMEN) that discriminates the summary-relevant information from a cluster of topic-related documents and generates a final summary via token probability marginalization.
- We advance the research in the medical domain, experimenting with a biomedical multi-document summarization dataset about the generation of systematic literature reviews.
- We show that our solution outperforms previous state-of-the-art solutions, achieving better ROUGE scores. Furthermore, we extensively prove the contribution of each module of our method with ablation studies.

2 Related Work

We describe related works on multi-document summarization categorized on model architectures.

Flat solutions. Flat concatenation is a simple yet powerful solution because the generation of the multi-document summary is treated as a single-document summarization task, thus it can leverage state-of-the-art pre-trained summarization models. Consequently, processing all documents as a flat input requires models capable of handling long sequences. As previously experimented by DeYoung et al. (2021), Xiao et al. (2021) proposed to leverage the Longformer-Encoder-Decoder model (Beltagy et al., 2020) pre-trained with a novel multi-document summarization specific task. They proved that a long-range Transformer that encodes all documents is a straightforward yet effective solution, and they achieved new state-of-the-art results in several multi-document summarization datasets. However, such models may struggle to handle a massive cluster of topic-related documents since they need to truncate them because of architectural limits. Further, processing all documents

in a cluster could be noisy if some of them are not relevant or factual with respect to the summary.

Hierarchical solutions. To better preserve cross-document relations and obtain semantic-rich representations, hierarchical concatenation solutions leverage graph-based techniques to work from word and sentence-level (Wan and Yang, 2006; Liao et al., 2018; Nayeem et al., 2018; Antognini and Faltings, 2019; Li et al., 2020) to document-level (Amplayo and Lapata, 2021). Other hierarchical approaches include multi-head pooling and inter-paragraph attention architectures (Liu and Lapata, 2019a), attention models with maximal marginal relevance (Fabbri et al., 2019), and attention across different granularity representations (Jin et al., 2020). Such models are often dataset-specific because of the custom architecture, so they struggle to adapt to other datasets and effectively leverage pre-trained state-of-the-art Transformers.

Our solution. In this work, we show how the summary-relevant information can be discriminated from a cluster of medical documents by a probabilistic neural method trained end-to-end. In detail, our solution fully leverages pre-trained state-of-the-art Transformers without applying input truncation that causes performance drop and discards important contents, unacceptable for a high-social impact domain such as the medical one.

3 Method

We introduce DAMEN, a discriminative marginalized probabilistic neural method for the multi-document summarization of medical literature based on three components:

- **Indexer:** it is a neural language model based on BERT architecture (Cohan et al., 2020) that creates a dense representation of documents in the cluster, according to the best practices for information retrieval systems.
- **Discriminator:** it leverages a BERT model to create the background embedding, which is used to compute a distance score between the embedding of each document in the cluster in order to select the top K ones.
- **Generator:** it uses a BART model (Lewis et al., 2020a) to produce the final summary via token probability marginalization from the top K documents combined with the background.

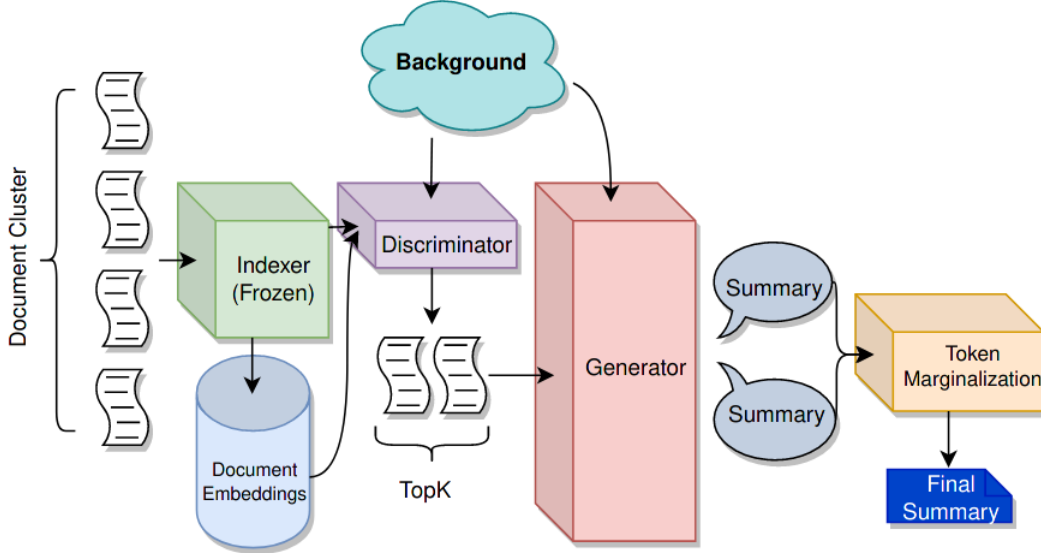


Figure 1: The overview of DAMEN, our probabilistic neural method for multi-document summarization of medical literature. First, two BERT models (i.e., *Indexer* and *Discriminator*) encode the background and documents in the cluster, creating dense embedding representations. Then, the *Discriminator* selects the top K documents via inner product with the background. Afterward, the background is concatenated with each document retrieved and the new textual inputs are given to BART to generate the multi-document summary by marginalizing the token probability distribution at decoding time.

While the *Indexer* is a frozen pre-trained model based on BERT, the *Discriminator* and *Generator* are trained end-to-end during the learning phase (Fig. 1). The overall task can be mathematically formalized as follows. The training tuple is composed of three elements (y_i, x_i, \mathbf{C}_i) , where y_i is the ground-truth target summary, \mathbf{C}_i is the cluster of documents used to generate the multi-document summary, and x_i is the background, which is a textual context shared by all $c_j \in \mathbf{C}_i$ used as input of the method, similar to the query in the query-focused multi-document summarization (Su et al., 2020). The whole pipeline is trained end-to-end to maximize the conditional probability of generating y_i from x_i and \mathbf{C}_i through gradient descent:

$$p(y_i|x_i, \mathbf{C}_i) \quad (1)$$

3.1 Indexer

In this phase, we index each document in the cluster with an embedding generated by a BERT-based model. Such a pre-trained language model is the state-of-the-art in semantic modeling from textual data thanks to the vast knowledge learned during pre-training (Chen et al., 2019), achieving ground-breaking results across an extensive range of NLP downstream tasks even without fine-tuning. For this reason, we use it to create a dense latent representation of each document, called document em-

bedding, which is a vector of continuous numbers that indicates a point in a latent semantic space. The technique we use is known as dense passage retriever (DPR) (Karpukhin et al., 2020), and it is widely adopted in the information retrieval domain (e.g., Lin et al., 2021; Moro and Valgimigli, 2021). We choose the DPR method because it does not interrupt the backpropagation, differently from other solutions, e.g., BM25, TF-IDF (Domeniconi et al., 2014b) or LSA (Domeniconi et al., 2016a).

We formalize this step as $B_\beta(\mathbf{C}_i) = E$, where B is a BERT-based model, β represents its parameters, and E is a matrix of shape $(len(\mathbf{C}_i), 768)$, where each row j of the matrix is the latent representation of the document c_j .

3.2 Discriminator

The main idea of the *Discriminator* is to discriminate the critical information from noise in a cluster of topic-related documents with respect to a shared background without breaking the backpropagation chain. For this reason, we use a probabilistic deep neural model to draw a probability distribution over documents in the cluster $\langle c_0, c_1, \dots, c_n \rangle \in \mathbf{C}_i$, with the following formula:

$$p_\theta(\mathbf{C}_i|x_i) \quad (2)$$

where θ represents the parameters of the neural network. Even in this case, the neural model is

Statistic	MS2		
	Background	Document	Summary
# average num. of tokens	125.75	546.90	74.52
# average num. of abstracts per background		23.30	
# backgrounds		13982	

Table 1: The dataset statistics. All values are mean over the whole dataset except for the “# backgrounds” row.

a BERT-based pre-trained language model as the one used for indexing, but this is trained during the learning process while the first is frozen. In detail, the *Discriminator* creates a latent projection for each background, which is used to fetch the more related documents in the cluster. More precisely, it applies the inner product to create a score for each document and selects the top K ones.

3.3 Generator

We use the pre-trained encoder-decoder generative Transformer BART (Lewis et al., 2020a) to summarize the C_i weighted by the *Discriminator*. This component is trained to predict the next output token, creating a probability distribution over the dictionary for each $c_j \in C_i$ before marginalizing. The process is then repeated for all the target tokens.

Before giving the documents to the model, we concatenate them with the background x_i , creating $c'_{ij} = [x_i, tok, c_{ij}]$, where *tok* is a special text separator token ($\langle doc \rangle$) we add between x_i and c_{ij} to make BART aware of the background text boundary. The behavior of the *Generator* can be formally defined as follows:

$$p(y_i | c'_{ij}) = \prod_z^N p_\gamma(y_{iz} | c'_{ij}, y_{i,1:z-1}) \quad (3)$$

where γ are the *Generator* parameters, $N = |y_i|$ is the target length, and $y_{i,1:z}$ are the tokens from position 1 to z of the target y_i .

3.4 Model

The entire model aims to draw the probability distribution over the dictionary to generate the output tokens y_i conditioned by x_i and C_i that we formally define as:

$$p(y_i | x_i, C_i) = \prod_z^N \sum_{c \in top-k} p_\theta(c_i | x_i) p_\gamma(y_{iz} | c_i, y_{i,1:z-1}) \quad (4)$$

We train the whole model by minimizing the negative marginal log-likelihood of each target with

the following loss:

$$-\sum \log p(y|x, C) \quad (5)$$

4 Experiments

This section starts with describing the dataset in §4.1 and training details in §4.2. We then analyze model performance in §4.3 and finally conduct ablation studies in §4.4.

4.1 Dataset

We tested and evaluated our proposed method on the only medical dataset for multi-document summarization, as far as we know, about the generation of systematic literature reviews: the MS2 dataset. The dataset is provided in DeYoung et al. (2021), and it is freely distributed.² It contains over 470K document abstracts and 20K summaries derived from the scientific literature. Each sample of the dataset is composed of three elements: i) the *background* statement, which is a short text that describes the research question or topic shared by all documents in the cluster, ii) the *target* statement, which is the multi-document summary to generate, and iii) the *studies*, also defined as cluster for consistency with our notation, which is a set of abstracts of medical studies related to the topic covered in the background statement.

The problem can be formalized as follows: we have a target statement to generate about the background source, containing the topic specifications, and a cluster of related document abstracts from which to fetch and discriminate helpful knowledge with respect to the background. From here on, we use the terms “document” and “abstract” interchangeably since the elements in the cluster are just the abstracts of medical documents.

We report the dataset statistics in Table 1.

²<https://github.com/allenai/ms2>

Model	MS2		
	Rouge-1	Rouge-2	Rouge-L
Prev. SOTA			
LED _{FLAT}	26.89	8.91	20.32
BART _{HIERARCHICAL}	27.56	9.40	20.80
Our			
DAMEN	28.95	9.72	21.83

Table 2: The results on MS2 for multi-document summarization of systematic literature reviews. The results of previous state-of-the-art are taken from DeYoung et al. (2021). Better ROUGE scores are bolded.

4.2 Training Details

We trained our solution for 3 epochs using a batch size of 1 and a learning rate with a linear schedule set to 1×10^{-5} . We set the number of K equal to 6 because it gave best results and used 1024 tokens as the max input size for the *Generator*. During the evaluation, we adopted a beam size of 4 with a min and a max length set to 32 and 256, respectively.

We implemented the code using PyTorch for tensor computations and Hugging Face³ for language model checkpoints. We performed the experiments on a workstation with a GPU Nvidia RTX 3090 of 24GB memory, 64GB of RAM, and a processor Intel(R) Core(TM) i9-10900X CPU @ 3.70GHz.

4.3 Results

Table 2 shows the results on multi-document summarization of systematic literature reviews, comparing our method with two solutions proposed in DeYoung et al. (2021). The BART_{HIERARCHICAL} solution is trained to encode each document independently and then concatenate the representation of hidden states before decoding, whereas LED_{FLAT} takes as input all documents concatenated as a single document. Experimental results show we outperform the state-of-the-art in all the ROUGE metrics, proving better capability to discriminate relevant information across many related documents and merge it consistently (Fig. 2).

4.4 Ablations

We conducted ablation studies on the MS2 dataset to prove the importance of each module of our method. In detail, for all experiments we trained our solution for 1 epoch with the same training details reported in §4.2, and we performed the evaluation on the first 400 instances of the test set.

³<https://huggingface.co/models>

The importance of a highly abstractive large-sized Generator. We report in Table 3 the performance using several pre-trained checkpoints of the *Generator* that differ in size and training. In detail, we tested two BART_{BASE} checkpoints and three BART_{LARGE} checkpoints:

- *facebook/bart-base*: the actual BART model pre-trained with a denoising masked language modeling.
- *gayanin/bart-mlm-pubmed*: the BART model pre-trained exclusively on scientific corpora.
- *facebook/bart-large*: the same BART model as the base version with a large architecture.
- *facebook/bart-large-cnn*: the large BART fine-tuned on single-document summarization on the CNN/DailyMail dataset (Nallapati et al., 2016).
- *facebook/bart-large-xsum*: the large BART fine-tuned on single-document summarization on the XSum dataset (Narayan et al., 2018).

Results prove that a large-sized BART model already fine-tuned on a summarization task achieves better performance. More precisely, the checkpoint fine-tuned on the XSum dataset obtains better results thanks to the higher abstractiveness and the shortness of the target summaries, which are made up of just 1-2 sentences, similar to the MS2 dataset.

The importance of a full-sized chunked representation of documents in the cluster. Table 4 reports experiments with three cluster configurations, where each document is treated with a different text representation, described as follows:

- *Document-level*: the simpler configuration that considers the entire abstracts in the cluster. We truncated documents taking only the first 512 tokens before encoding by the *Indexer*.
- *Sentence-level*: we considered the sentences of each document obtained using the state-of-the-art tokenizer PySBD (Sadvilkar and Neumann, 2020). The sentences are encoded up to 128 tokens in length and they are then treated as individual textual units.
- *Chunk-level*: our configuration, where each document is split into chunks of exact 512 tokens to consider all text information without

Background: An individual patient data meta analysis was performed to determine clinical outcomes, and to propose a **risk stratification** system, related to the comprehensive treatment of patients with **oligometastatic nsclc**.

Doc1: ... **We therefore did this phase iii trial to compare concurrent chemotherapy and radiotherapy** followed by resection with st and ard concurrent chemotherapy and definitive radiotherapy without resection ... **In an exploratory analysis, os was improved for patients who underwent lobectomy, but not pneumonectomy, versus chemotherapy plus radiotherapy. Chemotherapy plus radiotherapy** with or without resection (preferably lobectomy) are options for patients with stage iiia (n2) non-small-cell lung cancer.

Doc2: ... Common **adverse events** associated with crizotinib were visual disorder, gastrointestinal side effects, and elevated liver aminotransferase levels, whereas common adverse events with chemotherapy were fatigue, alopecia, and dyspnea. **Patients reported greater reductions in symptoms of lung cancer and greater improvement in global quality of life with crizotinib than with chemotherapy.**

Doc3: ... First-line gefitinib for patients with advanced non-small-cell lung cancer who were selected on the basis of egfr mutations **improved progression-free survival**, with acceptable toxicity, as **compared with st and ard chemotherapy** ...

Ground-truth: Significant os differences were observed in oligometastatic patients stratified according to type of metastatic presentation, and n status. Long-term survival is common in selected patients with metachronous oligometastases.

Model: The pooled **risk stratification** of patients with **oligometastatic nsclc** showed a **significant reduction in the risk of adverse events compared with st and ard chemotherapy, but not radiotherapy.**

Figure 2: A random sampled test set instance. We show how DAMEN selects the information from the background and multiple documents to generate the final summary.

Generator	MS2		
	Rouge-1	Rouge-2	Rouge-L
BART_{BASE}			
Original MLM	26.84	7.90	20.45
Scientific MLM	25.81	7.65	19.96
BART_{LARGE}			
Original MLM	24.81	7.21	19.31
CNN	27.22	8.30	20.99
XSum	28.35	8.96	21.62

Table 3: The ablations to validate the contribution of the *Generator*. Better ROUGE scores are bolded.

input truncation. This configuration is similar to the “sentence-level” one but with the difference that each textual unit is 512 tokens in length and not 128.

The results prove the better performance on a cluster with chunked documents. By considering 512 tokens for each document, we fully leverage the capability of BERT language modeling without truncating any information. Input truncation required by the “document-level” configuration plays an important role in final accuracy because it discards and ignores potential summary-relevant information, leading to a performance drop. The “sentence-level” setting lets us increase the top K sentences to retrieve, but it worsens the final summary because single sentences are too fine-grained.

Cluster	MS2		
	Rouge-1	Rouge-2	Rouge-L
Baselines			
Sentence-level	27.77	8.82	21.11
Document-level	28.14	8.82	21.33
Our			
Chunk-level	28.35	8.96	21.62

Table 4: The ablations to validate the best cluster configuration. Better ROUGE scores are bolded.

The importance of a background-first concatenation with special token. Table 5 reports the experiments with a different configuration of the concatenated inputs to give to the *Generator*. We experimented with four types of concatenation:

- *[Document + Background]*
- *[Background + Document]*
- *[Document + <doc> + Background]*
- *[Background + <doc> + Document]*

Results prove the importance of a background-first concatenation with the special token separator to make BART aware of the textual difference between the background and the documents.

The importance of pre-trained DPR encoders. Table 6 reports the experiments with different

Concatenation	MS2		
	Rouge-1	Rouge-2	Rouge-L
w/o token <doc>			
Document + Background	28.01	8.83	21.18
Background + Document	27.70	8.63	20.98
w/ token <doc>			
Document + Background	27.96	8.65	21.43
Background + Document	28.35	8.96	21.62

Table 5: The ablations to validate the input to give to the *Generator*. Better ROUGE scores are bolded.

model checkpoints for the *Indexer* and *Discriminator*. First, we leveraged the checkpoint “*sentence-transformers/allenai-specter*” (Cohan et al., 2020), which is a scientific BERT-based model trained to create document embeddings by using paper citations. Thus, we used this pre-trained model for both the *Indexer* and *Discriminator*. Second, we used two different checkpoints with a specific DPR training, such as “*facebook/dpr-question_encoder-single-nq-base*” for encoding the background and “*facebook/dpr-ctx_encoder-single-nq-base*” for encoding each document in the cluster.

Results prove the importance of the DPR checkpoints for both the *Indexer* and *Discriminator*.

5 Conclusion

We proposed a novel probabilistic method based on the combination of three language models to tackle multi-document summarization in the medical domain. This task is characterized by redundant information, noise, and the possible presence of vital information in each sentence that makes arbitrary input truncation unacceptable. For this reason, we proposed a multi-document summarization method able to discriminate salient contents from irrelevant before summarizing. In detail, the solution first leverages a BERT-based model (*Indexer*) for creating dense indices for each chunk of each document in the cluster. Then, a second BERT-based model (*Discriminator*) is used to process the shared background and select only the most relevant chunks. The final BART model is trained to perform a probability marginalization over each token prediction for each selected chunk. In this way, our solution reads all document information and selects just the most relevant chunks, discarding noise before feeding the *Generator*. The *Discriminator* and *Generator* are trained end-to-end, backpropagating the probability distribution as explained in §3. The *Indexer* is frozen; training would lead to some

Encoders	MS2		
	Rouge-1	Rouge-2	Rouge-L
w/o ad-hoc encoders			
SPECTER-based	27.79	8.60	21.23
w/ ad-hoc encoders			
DPR-based	28.35	8.96	21.62

Table 6: The ablations to validate the contribution of the *Indexer* and *Discriminator* model checkpoints. Better ROUGE scores are bolded.

problems, such as the time to learn improved embeddings at each iteration and the larger memory occupation to save the gradient for each document.

We tested our method on MS2, the only dataset on systematic literature reviews, and compared it with state-of-the-art models, finding that our novel approach outperforms competitors on the ROUGE evaluation metrics. Further, we performed extensive ablation studies to highlight the contribution of each component and motivate the design choices.

5.1 Future works

At the edge of our knowledge, this is the first work that applies a probability marginalization method for multi-document summarization. We believe this work can inspire novel research towards end-to-end multi-model collaboration instead of solutions with a single large model addressing the entire task. According to the *divide et impera* pattern, each model learns a specific sub-task, creating a more efficient and transparent cooperating solution. Tasks such as related work generation or text generation from multi-sourced inputs can get the most from our method, improving pre-existing solutions to discriminate helpful knowledge from noise.

Further possible directions to deal with multi-inputs are the following: i) extracting relevant snippets from documents with term weighting techniques (Domeniconi et al., 2015) or semantic relations with unsupervised methods (Domeniconi et al., 2016b, 2017) to better model interpretable representations based on knowledge graph learning techniques (Frisoni and Moro, 2020; Chen et al., 2021a,b) or event extraction methods (Frisoni et al., 2021); ii) training models to write and read cross-document information with self-supervised representation learning methods (Domeniconi et al., 2014c) and memory-based neural layers (Moro et al., 2018; Cui and Hu, 2021).

6 Ethical Considerations

The advancement of deep neural network architectures and the availability of large pre-trained language models has led to significant improvements for the multi-document summarization task, which has applications in high-impact domains, particularly in the medical one. Here, systematic literature reviews play an essential role for the medical and scientific community, and for that reason, they require strong guarantees about the factuality of the output summary. Current state-of-the-art NLP solutions cannot establish such assurance, so we do not believe our solution, like previous ones, is ready to be deployed. The research should explore more effective evaluation measures for text summarization to make it happen, and large-scale accuracy guarantees by medical experts are still needed. Finally, if the method will be applied to sensitive data such as medical patient records, it should also include privacy-preserving policies (da Silva et al., 2006).

Acknowledgments

We thank the Maggioli Group⁴ for granting the Ph.D. scholarship to L. Ragazzi and L. Valgimigli. The solution presented in this work has been designed by G. Moro.

References

Reinald Kim Amplayo and Mirella Lapata. 2021. [Informative and controllable opinion summarization](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2662–2672. Association for Computational Linguistics.

Diego Antognini and Boi Faltings. 2019. [Learning to create sentence semantic relation graphs for multi-document summarization](#). *CoRR*, abs/1909.12231.

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

Moye Chen, Wei Li, Jiachen Liu, Xinyan Xiao, Hua Wu, and Haifeng Wang. 2021a. [SgSum:transforming multi-document summarization into sub-graph selection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4063–4074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

⁴In particular, Manlio Maggioli, Paolo Maggioli, Cristina Maggioli, Amalia Maggioli, Nicoletta Belardinelli and Andrea Montefiori. <https://www.maggioli.com/who-we-are/company-profile>

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xi-angliang Zhang, Dongyan Zhao, and Rui Yan. 2021b. [Capturing relations between scientific papers: An abstractive model for related work section generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6068–6077. Association for Computational Linguistics.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2019. [Distilling the knowledge of BERT for text generation](#). *CoRR*, abs/1911.03829.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. [SPECTER: document-level representation learning using citation-informed transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2270–2282. Association for Computational Linguistics.

Peng Cui and Le Hu. 2021. [Sliding selector network with dynamic memory for extractive summarization of long documents](#). In *NAACL-HLT 2021, June 6-11*, pages 5881–5891. ACL.

Josenildo Costa da Silva, Matthias Klusch, Stefano Lodi, and Gianluca Moro. 2006. [Privacy-preserving agent-based distributed data clustering](#). *Web Intell. Agent Syst.*, 4(2):221–238.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. [MS²: Multi-document summarization of medical studies](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pietro di Lena, Giacomo Domeniconi, Luciano Margara, and Gianluca Moro. 2015. [GOTA: GO term annotation of biomedical literature](#). *BMC Bioinform.*, 16:346:1–346:13.

Giacomo Domeniconi, Marco Masseroli, Gianluca Moro, and Pietro Pinoli. 2014a. [Discovering new gene functionalities from random perturbations of known gene ontological annotations](#). In *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy, 21 - 24 October, 2014*, pages 107–116. SciTePress.

Giacomo Domeniconi, Gianluca Moro, Andrea Pagliarani, Karin Pasini, and Roberto Pasolini. 2016a. [Job recommendation from semantic similarity of linkedin users’ skills](#). In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2016, Rome, Italy, February 24-26, 2016*, pages 270–277. SciTePress.

- Giacomo Domeniconi, Gianluca Moro, Andrea Pagliarani, and Roberto Pasolini. 2017. [On Deep Learning in Cross-Domain Sentiment Classification](#). In *IC3K 2017*, volume 1, pages 50–60. SciTePress.
- Giacomo Domeniconi, Gianluca Moro, Roberto Pasolini, and Claudio Sartori. 2014b. [Cross-domain text classification through iterative refining of target categories representations](#). In *KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Rome, Italy, 21 - 24 October, 2014*, pages 31–42. SciTePress.
- Giacomo Domeniconi, Gianluca Moro, Roberto Pasolini, and Claudio Sartori. 2014c. [Iterative refining of category profiles for nearest centroid cross-domain text classification](#). In *Knowledge Discovery, Knowledge Engineering and Knowledge Management - 6th International Joint Conference, IC3K 2014, Rome, Italy, October 21-24, 2014, Revised Selected Papers*, volume 553 of *Communications in Computer and Information Science*, pages 50–67. Springer.
- Giacomo Domeniconi, Gianluca Moro, Roberto Pasolini, and Claudio Sartori. 2015. [A Comparison of Term Weighting Schemes for Text Classification and Sentiment Analysis with a Supervised Variant of tf.idf](#). In *DATA (Revised Selected Papers)*, volume 584, pages 39–58. Springer.
- Giacomo Domeniconi, Konstantinos Semertzidis, Vanessa López, Elizabeth M. Daly, Spyros Kotoulas, and Gianluca Moro. 2016b. [A novel method for unsupervised and supervised conversational message thread detection](#). In *DATA 2016 - Proceedings of 5th International Conference on Data Management Technologies and Applications, Lisbon, Portugal, 24-26 July, 2016*, pages 43–54. SciTePress.
- Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. [Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1074–1084. Association for Computational Linguistics.
- Giacomo Frisoni and Gianluca Moro. 2020. [Phenomena explanation from text: Unsupervised learning of interpretable and statistically significant knowledge](#). In *Data Management Technologies and Applications - 9th International Conference, DATA 2020, Virtual Event, July 7-9, 2020, Revised Selected Papers*, volume 1446 of *Communications in Computer and Information Science*, pages 293–318. Springer.
- Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. 2021. [A survey on event extraction for natural language understanding: Riding the biomedical literature wave](#). *IEEE Access*, 9:160721–160757.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. [REALM: retrieval-augmented language model pre-training](#). *CoRR*, abs/2002.08909.
- Thomas Hofmann. 1999. [Probabilistic latent semantic analysis](#). In *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pages 289–296. Morgan Kaufmann.
- Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. [Multi-granularity interaction network for extractive and abstractive multi-document summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6244–6254. 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.
- Khalid S Khan, Regina Kunz, Jos Kleijnen, and Gerd Antes. 2003. Five steps to conducting a systematic review. *Journal of the royal society of medicine*, 96(3):118–121.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. [BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. [Leveraging graph to improve abstractive multi-document summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6232–6243. Association for Computational Linguistics.
- Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. [Abstract meaning representation for multi-document summarization](#). In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1178–1190. Association for Computational Linguistics.
- Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. [Pretrained transformers for text ranking: Bert and](#)

- beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325.
- Yang Liu and Mirella Lapata. 2019a. [Hierarchical transformers for multi-document summarization](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5070–5081. Association for Computational Linguistics.
- Yang Liu and Mirella Lapata. 2019b. [Text summarization with pretrained encoders](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics.
- Gianluca Moro, Andrea Pagliarani, Roberto Pasolini, and Claudio Sartori. 2018. [Cross-domain & In-domain Sentiment Analysis with Memory-based Deep Neural Networks](#). In *IC3K 2018*, volume 1, pages 127–138. SciTePress.
- Gianluca Moro and Luca Ragazzi. 2022. [Semantic Self-Segmentation for Abstractive Summarization of Long Legal Documents in Low-Resource Regimes](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Virtual Event, February 22 - March 1, 2022*, pages 1–9. AAAI Press.
- Gianluca Moro and Lorenzo Valgimigli. 2021. [Efficient self-supervised metric information retrieval: A bibliography based method applied to COVID literature](#). *Sensors*, 21(19):6430.
- Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence rnns and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.
- Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Ylias Chali. 2018. [Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion](#). In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1191–1204. Association for Computational Linguistics.
- Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. [Efficiently summarizing text and graph encodings of multi-document clusters](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4768–4779. Association for Computational Linguistics.
- Dragomir R. Radev. 2000. [A common theory of information fusion from multiple text sources step one: Cross-document structure](#). In *Proceedings of the SIGDIAL 2000 Workshop, The 1st Annual Meeting of the Special Interest Group on Discourse and Dialogue, 7-8 October 2000, Hong Kong*, pages 74–83. The Association for Computer Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Nipun Sadvilkar and Mark Neumann. 2020. [PySBD: Pragmatic sentence boundary disambiguation](#). In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.
- Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J. Barezi, and Pascale Fung. 2020. [Cairo-covid: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management](#). In *EMNLP 2020, Online, December 2020*. ACL.
- Xiaojun Wan and Jianwu Yang. 2006. [Improved affinity graph based multi-document summarization](#). In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA*. The Association for Computational Linguistics.
- Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. [PRIMER: pyramid-based masked sentence pre-training for multi-document summarization](#). *CoRR*, abs/2110.08499.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. [Big bird: Transformers for longer sequences](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.