Leveraging Entity Pyramid-based Masked Sentence Pre-training and Graph Encodings for Multi-document Abstractive Summarization of Medical Literature Reviews

Rishi Dinesh[1], Dr. Priyadarshini J[1]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, 600127

Abstract

Systematic literature reviews seek to thoroughly summarize the data from all studies that are available and pertinent to a research question. Such evaluations represent the best available evidence in medicine that is used to guide clinical care. However, manual evaluations are expensive to conduct, taking teams of specialists months to complete, and soon become outdated. New, potent text summarizing methods are required to process this volume of text data quickly and effectively. The objective is to create a model that can provide a summary that synthesizes the data from the input studies given the abstracts of input studies relevant to a research question and a background section outlining that research issue. To this extent, this paper proposes a method that leverages graph encodings by representing each input study's abstract as one or more IDEO (Intervention, Direction, Evidence, Outcome) tuples and integrating this into the PRIMERA (Pyramid-based Masked Sentence Pre-training for Multi-document Summarization) transformer model. Since a "good" machine-generated summary must effectively capture cross-document relationships, adding linearized IDEO text as graph embeddings and applying PRIMERA's global attention to these tokens will enable the model to synthesize a summary that highlights the key information from all studies. Results of experiments show that the proposed model, trained in a 100-shot setting, outperforms existing benchmarks in the Multi-document Summarization for Literature Review (MSLR) shared task, proposed by the Allen Institute for AI (AI2). In terms of the Evidence Inference (ΔEI) metrics, the model achieves an ΔEI-Avg of 0.435 and an ΔEI-F1 of 0.285. The ΔEI-F1 metric shows a 16.9% improvement over the current state-of-the-art in this metric, highlighting the model's effectiveness. In a fine-tuned setting, the model yields a R-1 score of 0.227, R-2 score of 0.048, R-L score of 0.169 and BERTScore of 0.849, which is comparable to the benchmarks of the MSLR shared task.

Keywords: Multi-Document Summarization, Abstractive Summarization, Natural Language Processing, Transformers

1. Introduction

Systematic literature reviews are key elements of evidence based healthcare in the biomedical domain as they summarize medical knowledge pertaining to a specific issue by clearly formulating the research question, collecting quality relevant studies, and following explicit methodologies to arrive at a conclusion (Khan et al., 2003). These reviews, however, are an expensive and time-consuming process (Michelson & Reuter, 2019), each taking around 1-2 years to complete. The recent advancements in the field of deep learning in general, and Natural Language Processing (NLP) in specific, call for automated or semi-automated approaches to summarize salient information from multiple literatures with accuracy and efficiency. This is formulated as a Multi-Document Summarization (MDS) task in the NLP domain. MDS is

the task of creating a concise and comprehensive summary of multiple input texts written on the same topic; and this can be done either through an extractive, abstractive or hybrid approach.

MDS is more challenging than its single document counterpart, Single Document Summarization (SDS), on account that it must capture cross-document relations in a larger search space and often deal with contradicting and redundant information. This is especially true in the biomedical domain, where several clinical studies may provide contradicting and redundant evidence pertaining to a particular intervention. While several approaches have been proposed (refer section 2) that summarize Wikipedia (P. J. Liu et al., 2018) or news articles (Fabbri et al., 2019) , very little has been done in the biomedical domain. In order to promote research in this field, the Allen Institute for Artificial Intelligence proposed the Multi-document Summarization for Literature Reviews (MSLR) shared task using two high-quality biomedical MDS datasets, namely the Cochrane dataset (Wallace et al., 2021) and the Multi-Document Summarization of Medical Studies (MS2) dataset (DeYoung et al., 2021, p. 2). The scope of implementation in this paper is limited to the MS2 dataset, which consists of 20K reviews and 470k studies summarized by these reviews. In addition to supporting the MDS task by providing document-summary pairs, this dataset also provides a structured form of the data by using existing biomedical information extraction systems (DeYoung et al., 2020; B. Nye et al., 2018), which will also be leveraged as the graph encodings in the proposed method of this paper.

The aim of the MSLR task is to develop a model that can create a summary that captures the salient information from the input studies given the abstracts of the input studies relevant to a research question and a background section outlining that research topic. This synthesized summary should indicate an overall evidence direction (increase, decrease or no change) of an outcome for an intervention when measured on a population.

The rest of this paper is organized as follows: Section 2 covers some of the various techniques used in the field of MDS and highlights their drawbacks, thereby justifying the reason behind adopting the proposed technique. Following this, Section 3 describes the dataset and presents a brief Exploratory Data Analysis (EDA). Next, Section 4 introduces the proposed graph representation of the data and details the tokenization process. Then, Section 5 explains the proposed PRIMERA method and how it is integrated with the graph representation. Consequently, Section 6 describes the experimental setup whereas Section 7 lays out the results of this paper. Finally, Section 8 highlights the various limitations and scope for future work and Section 9 provides the conclusion.

2. Related Works

Recent work in the field of abstractive summarization has used deep learning based models that are based on one or more of the following techniques: Recurrent Neural Networks (RNN) and its variants such as Gated Recurrent Units (GRU), Long Short-term Memory (LSTM) and Bidirectional LSTMs (Bi-LSTM), Convolutional Neural Networks (CNN), Graph Neural Networks (GNN), Pointer-Generator (PG) networks and Transformers (Ma et al., 2022). Since the MSLR shared task is new at the time of writing this paper, there has not been much work on it[1]. This section thus surveys abstractive summarization approaches in

[1] The models submitted for the MSLR shared task will be referenced in Section 7.3 as baselines for comparison.

other domains with the hope to translate it or at least derive inspiration for the biomedical domain. This survey is largely derived from the works of (Ma et al., 2022).

2.1 Abstractive Summarization

Several approaches have used RNNs for the task of MDS, such as in (Zhang et al., 2018), that used a hierarchical LSTM based encoder-decoder framework with an attention mechanism to extend a SDS model for MDS. Similarly, (Wang & Ling, 2016) also adopted an attention based encoder-decoder architecture based on LSTMs for abstractive summarization of opinionated text. (Amplayo & Lapata, 2021) modelled the task of opinion summarization using a two-stage Condense-Abstract (CA) framework, wherein the condense model used a bi-LSTM auto-encoder to condense the input reviews into dense vectors, which was then fed as input into the abstract model that used a vanilla LSTM enhanced with attention and copy mechanisms to generate the summary. (Bražinskas et al., 2020a), on the other hand, proposed a GRU based Variational Auto-Encoder (VAE) architecture for abstractive opinion summarization in an unsupervised setting.

Many other unsupervised based approaches using RNNs were proposed for MDS due to the lack of availability of high-quality document-summary pairs. For instance, (P. Li et al., 2017) proposed an RNN based encoder-decoder framework with a cascaded attention model that was used to estimate the salience of words and sentences , which was then used for compressive summary generation. Along the same lines, (Chu & Liu, 2019) proposed an end-to-end model architecture for abstractive summarization using an RNN-based auto-encoder where the mean of the representations of the input reviews decodes to a reasonable summary-review while not relying on any review-specific features. Building on this, (Coavoux et al., 2019) used an LSTM to encode input representations, clustering and aggregating to group similar representations together, and generating a sentence for each cluster.

Research has shown that CNNs can be used to learn a sequence of feature representations and that hierarchical CNN structures can extract representations that benefit an abstractive summarization task (Gehring et al., 2017). The convolutional kernels' parameter sharing allows the model to extract specific types of features such as n-gram features. This was leveraged by (Lin et al., 2018) that added a convolutional gated unit with self-attention to the encoder of a seq2seq model for global encoding. Several other papers have also integrated a CNN layer into the encoder-decoder architecture of a seq2seq model for better performance (Song et al., 2019; C. Yuan et al., 2020). (Y. Liu et al., 2018) even extended the CNN seq2seq model proposed by (Gehring et al., 2017) to generate abstractive summaries of desired lengths.

Several papers have also adopted GNNs for the task of summarization due to their ability to model the semantic and syntactic relationships between entities in natural language. The two popular common forms of the GNN are the Graph Convolutional Network (GCN) (Kipf & Welling, 2017) and the Graph Attention Network (GAN/GAT) (Veličković et al., 2018). GCNs are used to apply convolution on non-Euclidean document structures. (Q. Yuan et al., 2021) integrated GCN layers into an encoder-decoder architecture that concatenate node-level and graph-level representations of word embeddings to the seq2seq model whereas (Xu et al., 2020) built a syntactic document encoder that applies stacked layers of GCN on a document-level graph constructed using dependency parsing trees. On the other hand, GATs use an attention method to work with graph-structured data and computes each node's hidden

representations by attending to those of its neighbors. (Jin et al., 2020a) integrated a graph encoder built using GAT layers into a transformer-inspired encoder-decoder attention framework.

In order to alleviate the issue of factual errors and high redundancy in summarization tasks, PG networks were proposed (See et al., 2017). (Fabbri et al., 2019) integrated the Maximal Marginal Relevance (MMR) method into a hierarchical PG network for MDS. This model improved PG networks by using the MMR technique, which considers sentence *importance* as well as *redundancy*, to modify the attention weights for summary generation. Similarly, (Lebanoff et al., 2018) also proposed an integration of MMR with PG networks, but followed a different approach in calculating the MMR scores.

Transformer based architectures (Vaswani et al., 2017) can overcome the issues of vanishing gradients and information bottlenecks associated with processing long sequences while allowing for parallelization using a self-attention mechanism and have proven successful in several MDS tasks. For instance, (P. J. Liu et al., 2018) used extractive summarization to identify salient information and passed it into transformer-based decoder-only architecture to generate an abstractive summary. (Jin et al., 2020b) model words, sentences and documents as different granularities of semantic units using a hierarchical relation graph and use self-attention and cross-attention mechanisms to unify extractive and abstractive MDS. (Bražinskas et al., 2020b) leverage few-shot learning using a transformer conditional language model for both extractive and abstractive summarization.

Several approaches have also captured the hierarchical structures and relations that might exist among documents using graphs and hierarchical transformers. (Y. Liu & Lapata, 2019) augmented the method proposed by (P. J. Liu et al., 2018) using an inter-paragraph and graph-informed attention mechanism that enabled the model to encode documents hierarchically. (W. Li et al., 2020) also leveraged graph encodings of documents to capture cross-document relations and combined them with pre-trained language models for MDS. (Pasunuru et al., 2021) presented a graph enhanced approach to MDS using a pre-trained encoder-decoder transformer (Lewis et al., 2019) equipped with an efficient encoding mechanism to handle long inputs (Beltagy et al., 2020).

Pre-trained Language Models (LM) have shown great success in downstream tasks such as abstractive MDS. As shown in (W. Li et al., 2020), replacing low-level encoding layers with pre-trained LMs can improve performance. The approach used by (Pasunuru et al., 2021) can be integrated with pre-trained LMs to overcome the quadratic complexity of self-attention mechanisms. Pre-trained LMs such as BART (Lewis et al., 2019), GPT (Radford et al., n.d.) and T5 (Raffel et al., n.d.) have been successfully used in MDS tasks (Goodwin et al., 2020). Another transformer model, PEGASUS (Zhang et al., 2020), was pre-trained with Gap Sentence Generation (GSG) that focused on SDS. PRIMERA (Xiao et al., 2022), on the other hand, is a pre-trained LM specifically designed for MDS.

From this survey, it was observed that while RNN based auto-encoders have proven widely successful in SDS tasks, they fail to encode salient information in MDS tasks due to information bottlenecks. Even CNN and GNN based hybrid seq2seq models suffer from this limitation. Refer to Table 5 of the Appendix for a detailed comparative analysis of the above-mentioned deep learning approaches for abstractive summarization.

While transformers can overcome most of the drawbacks of auto-encoders using different attention mechanisms, majority of state-of-the-art architectures use a Masked Language Modelling (MLM) pre-training objective, which when applied on downstream tasks like MDS, could limit their performance

(Zhang et al., 2020). Thus, this paper uses the PRIMERA model, which was pre-trained using a GSG objective, and whose architecture is the same as that of the state-of-the-art Longformer Encoder-Decoder (LED) model.

3. Dataset Description

3.1 The MS2 (Multi-Document Summarization of Medical Studies) Dataset

As mentioned in Section 1, the MS2 dataset consists of 470k documents summarized by 20k systematic literature reviews. This dataset is constructed from papers in the Semantic Scholar literature corpus filtered using a suitability criterion. Each data item is a review represented as a JSONL object with a title, abstract and several studies, each with their own titles and abstracts (among other attributes). Moreover, each sentence of a given review abstract is classified as either *background* or *target* (a total of 9 class labels were considered but only these 2 are of interest). Since reviews that address similar research questions might cite the same studies, reviews are first clustered together before splitting into train, test and validation sets to prevent the models from learning from the test data.

In addition to providing document-summary pairs, this dataset also provides a structured representation of the same to give more details for interpreting the conclusions reported in each paper as well as to help evaluate the factual consistency between the input studies and reviews. To this extent, each study and review abstract is also tagged with PICO[2] (Population, Intervention, Comparator and Outcome) elements, as it is known to be a common way of representing clinical knowledge (Huang et al., 2006). However, recent work (B. E. Nye et al., 2020) has shown that the Comparator can be removed from this set without any hindrance to performance

Furthermore, every I/O (Intervention-Outcome) pair found in the review abstract is also associated with an evidence direction D (which can be classified into one of *increases, no change, decreases*) and an evidence sentence E supporting the evidence direction classification. This was done using the Evidence Inference (EI) dataset and models proposed by (DeYoung et al., 2020). The same set of I/O pairs found in the review abstract was used to generate evidence directions D and evidence sentences E in the studies. Thus, each study can be represented as a set of one or more IDEO (Intervention, Direction, Evidence, Outcome) tuples and a review, being a collection of studies, could also be represented as the same. This will be used as the "graph representation" of the input as described in Section 4.2.

3.2 Exploratory Data Analysis

The MSLR shared task provided a cleaned version of the MS2 dataset through the huggingface library, which is used in this paper. Each instance in this dataset contains the review ID, the review abstract, and the titles, abstracts and PMID of all the studies included in the review. After filtering out reviews for which

[2] Any clinical question in evidence-based medicine can be formulated in terms of the Population P (*who was studied?),* Intervention I (*what intervention was studied?*), Comparison C (*what was the intervention compared against?*) and Outcome O (*what was measured?*).

no IDEO pairs could be generated (explanation in Section **Error! Reference source not found.**), the training, validation and test set consisted of 4493, 625 and 534 instances respectively. With this, it was observed that each review included roughly 27 studies on average, with 75% of the reviews containing only 30-34 studies. Moreover, it was found that each study, on average, could be represented with 3 IDEO tuples, with 75% of the studies being represented with just 4 IDEO tuples. Figure 1 shows the count distribution of studies per review and IDEOs per study. Refer to Table 6 of the Appendix for complete statistics.

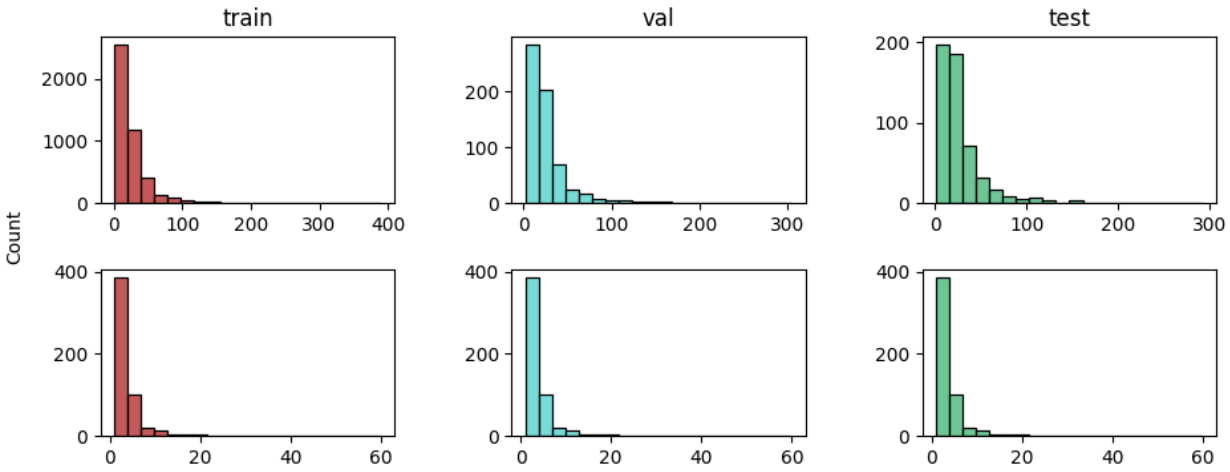

Figure 1 Count histogram of studies per review (top) and IDEOs per study (bottom) for each subset

4. Data Preparation

4.1 Graph Representation

Along with the MS2 dataset, the authors of (DeYoung et al., 2021, p. 2) also proposed two tasks, one *text-to-text* task to generate summaries given study abstracts as input, and another *table-to-table* task, to generate an evidence direction, given a structured representation of the input. This paper effectively combines the two forms of input representations (a form of *text-table to text* task) to better capture the cross-document relationships that exists between the documents. Each study (and each review) can be visualized as a graph with Interventions and Outcomes as nodes and the Evidence-Direction as the connecting edges (the direction can be thought of as an edge weight). This graph representation can succinctly capture the essence of the review document and when combined with other review graphs, can easily highlight the cross-document relationships that exist between them.

It was found that resorting to using the full cross-product of Interventions and Outcomes resulted in duplicated I/O pairs as well as potentially spurious pairs that do not correspond to actual I/O pairs in the review (DeYoung et al., 2021, p. 2). Thus, for each review, only the I/O pairs that are not in the review abstract's *effect*[3] statement was considered since these sentences formed the target summary for the

[3] As mentioned in Section 3.1, each sentence of the review abstract is classified into one of 9 class labels; *effect* is one of those labels.

review. Due to this, a fraction of the data instances ended up with no IDEO tuples for its studies and was thus filtered out of the dataset.

For example, consider the sample review abstract given in Table 1, where each row is a sentence of the abstract. The PICO elements in the text are tagged using special tokens: <pop>, </pop>, <int>, </int>, <out>, and </out>. Since sentences 3 and 4 have been tagged with EFFECT, they are removed from consideration along with the associated interventions and outcomes. Thus, the interventions considered are *prebiotics* and *probiotics*, whereas the only outcome considered is *efficacy.* Therefore, each study abstract used in this review will be represented as a set of two IDEO tuples, one for *probiotics-efficacy*, and another for *prebiotics-efficacy*.

Table 1 Sample review abstract tagged with sentence labels and PICO spans

S. No	Label	Sentence	Interventions	Outcomes
1.	BACKGROUND	Necrotizing enterocolitis (NEC) is one of the most serious gastrointestinal emergencies in <pop> very low birth weight (VLBW) preterm neonates </pop>, affecting 7 - 14 % of these neonates.	-	-
2.	BACKGROUND	Due to the seriousness of the disease, prevention of NEC is the most important goal.	-	-
3.	EFFECT	Current evidence from systematic review and meta- analysis revealed that **<int> probiotics </int>** are the most promising intervention in reduction of the incidence of **<out> NEC </out>** in <pop> VLBW neonates </pop>	probiotics	NEC
4.	EFFECT	As per the evidence, **<int> prebiotics </int>** modulate the composition of human intestine microflora to the benefit of the host by **<out> suppression </out>** of **<out> colonization of harmful microorganism </out>** and /or **<out> the stimulation of bifidobacterial growth, decreased stool viscosity, reduced gastrointestinal transit time and better feed tolerance. </out>**	prebiotics	Suppression Colonization of harmful microorganism the stimulation of bifidobacterial growth, decreased stool viscosity, reduced gastrointestinal transit time and better feed tolerance.
5.	ETC	**<int> Prebiotics </int>** may be potential alternatives or adjunctive therapies to **<int> probiotics </int>,** despite a lack of evidence supporting its clinical **<out> efficacy </out>** in prevention of NEC.	prebiotics probiotics	Efficacy
6.	ETC	In this article, we discuss evidence -based physiological effects of **<int> prebiotics </int>** and its therapeutic role in prevention of NEC.	prebiotics	-

4.2 Tokenization

Since PRIMERA uses the LED model as its backbone, the tokenizer used is the *LEDTokenizer* from the huggingface Tokenizer library. The corresponding start and end tokens for the IDEO tuples such as <int>, </int>, <direction>, </direction>, <evidence>, </evidence>, <out> and </out> were added as additional special tokens to the tokenizer. This tokenizer for PRIMERA also has an extra <doc-sep> special token to separate input documents.

To estimate the number of studies per review to include as part of the input to the model, a simple EDA was performed on the encoded abstract and IDEO lengths. Figure 2 shows a distribution of the abstract and IDEO encoding lengths per review. Specifically, for each review, the abstracts and IDEOs of each study was encoded, and the average of their lengths was considered as the abstract and IDEO encoded length of that review. It was observed that on average, the encoded abstract length of a review was roughly 377 tokens long, whereas the average encoded IDEO length of a review was around 210 tokens long. Combining these two, the average encoded review length would be equal to 587 tokens, which when rounded to the nearest power of 2 is 512 tokens. Thus, each input document was set to be truncated at a length 512 tokens during tokenization, with the first 256 tokens used for abstract text encoding and the next 256 tokens used for IDEO encoding. Since PRIMERA's maximum input length is 4096 tokens, the number of studies to include per review was set to be 4096/512 = 8 studies per review. Refer to Table 7 of the Appendix for complete statistics on the abstract and IDEO encoding lengths.

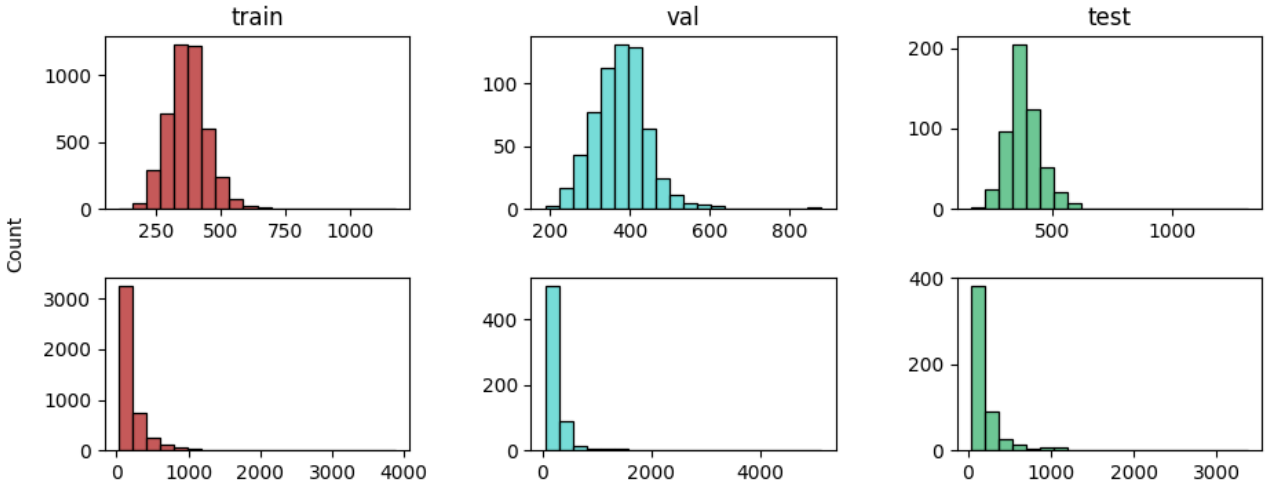

Figure 2 Count histogram of average abstract encoding lengths (top) and IDEO encoding lengths (bottom) per review per subset

5. Proposed Method

5.1 PRIMERA

PRIMERA (Pyramid-based Masked Sentence Pre-training for Multi-document Summarization) is a generation model pre-trained specifically for multi-document summarization in a way that reduces the dependency on large, fine-tuned data and dataset specific architectures, thereby enabling it to perform well in zero-shot and few-shot settings. The model's architecture is based on the Longformer Encoder-Decoder (LED) model introduced by (Beltagy et al., 2020), which allows it to efficiently handle long inputs using sparse local + global attention mechanism in the encoder stack. It takes in a flat concatenation of the documents as input separated by a special token to which the global attention is assigned to make the model aware of the document boundaries.

During pretraining, PRIMERA uses a strategy that is intended to teach the model how to recognize and group important information from a "cluster" of related documents. Specifically, It builds on the Gap Sentence Generation (GSG) pre-training objective introduced by the PEGASUS model (Zhang et al., 2020). GSG is the process of masking out several sentences in the input document and training the model to generate them. PRIMERA uses the same objective, but proposes a different strategy to select sentences for masking, called *Entity Pyramid Masking,* which was inspired by (Nenkova & Passonneau, 2004). This strategy uses the frequency of named entities in a cluster of documents to select salient sentences for masking. This results in masking sentences that are representative of more documents in the cluster than the being exact matches between fewer documents.

5.2 PRIMERA with IDEO attention

PRIMERA uses longformer's sparse windowed local attention along with a task-motivated global attention in its encoder to overcome the quadratic complexity of standard attention mechanisms. The local attention uses a fixed size sliding window around each token such that each token attends only to all other tokens within the window. By stacking multiple layers of such windowed attention, a huge receptive field is created, where the top layers have access to all input locations and can create representations that include data from all the input.

The global attention in the standard PRIMERA model on the other hand is added to the <doc-sep> token separating the documents in the input. This makes the model aware of the document boundaries and can also be used to share information across documents (Caciularu et al., 2021). In addition to this, this paper proposes adding global attention to the IDEO entity spans to make the model explicitly attend to the condensed information brought out by the graph encodings. In other words, local attention is added to the text representation of the input whereas global attention is added to the graph representation.

This approach was inspired by the work proposed by (Otmakhova et al., 2022), who also used a PRIMERA model on the Cochrane dataset for the MSLR shared task but placed global attention on PICO spans. Using the experimental results of (Otmakhova et al., 2022), this paper assigns global attention only to the tokens inside the opening (<>) and closing (</>) IDEO tags as this has shown to yield the highest scores. The IDEO tags themselves are replaced with padding tokens to mask them in the inputs and thus do not get either local or global attention.

Figure 3 depicts a sample review being passed to the core LED model; all the studies of the review are concatenated together (separated by the <doc-sep> token), with each study's abstract being appended with its corresponding graph representation.

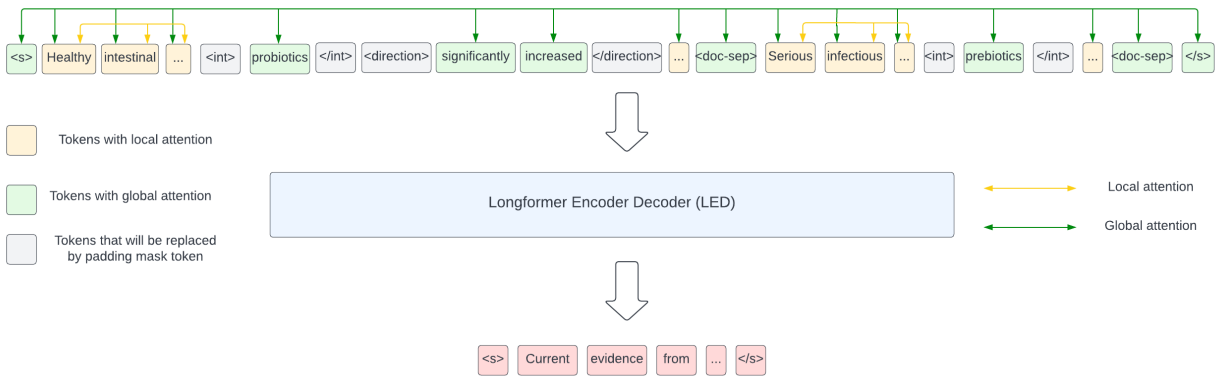

Figure 3 Overview of proposed method

6. Experimental Setup

The proposed approach was experimented in zero-shot, few-shot (10-shot, 100-shot) and fine-tuned settings to test the utility of the pre-training strategy as well as the impact of the addition of graph encodings. Since the results of the few-shot setting is dependent on the nature of the data sampled, each few-shot experiment was run five times with different random seeds and the one with the best results was chosen. A hyperparameter tuning sweep was performed for some of the parameters such as the number of beams (1, 2 and 3), length penalty (1 and 2), maximum output length (128 and 256) as well as the number of studies (8, 16 and 24). The setting with 1 beam, a length penalty of 1 and maximum output length of 128 for inputs with 8 studies per review yielded the highest average ROUGE score, which was consistent with the hyperparameters reported in the original PRIMERA paper (Xiao et al., 2022) and also validated the estimation provided in Section 4.2. Other hyperparameters were the same as that used in the original paper.

To evaluate the generated summaries, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores and BERT scores are employed. The ROUGE-N (R-N) score measures the number of matching N-grams between the generated and reference summary. The BERT score, on the other hand, is used to compute the semantic similarity between the tokens of the target and the prediction. The R-1, R-2 and R-L[4] F1-scores along with the average BERT F1-score is reported in this paper. In addition to this, this paper also reports the average and macro F1 Evidence-Inference (ΔEI) based divergence metric introduced in (DeYoung et al., 2021, p. 2). This metric captures the agreement of Interventions, Outcomes and EI directions between the input studies of the review and the generated summary.

Specifically, as described in Section 3.1, each I/O pair can be associated with an evidence direction classification, which is a probability distribution over three classes – *increases, no change, decreases*. The ΔEI metric computes the average Jensen-Shannon Distance (JSD) of these probability distributions between the predicted and target summaries for all I/O tuples in the review. This metric was introduced

[4] The R-L score measure the longest common subsequence between the prediction and the target.

as an attempt to overcome the weakness of the ROUGE scores in capturing the factuality of the generated summaries and was also a driving factor for adopting the proposed approach; the authors believed that by training PRIMERA with global attention on IDEO representation of the input, the model would be able to better generate summaries with a lower EI divergence score.

7. Results and Discussion

7.1 Experimental Results

Table 2 shows the results of the proposed approach; as expected, the fine-tuned version of the model yielded the highest ROUGE and BERT scores. However, the model trained in a 100-shot setting had the best ΔEI metrics, outperforming even the current benchmarks (refer Table 4). With a difference of less than 0.02 in average ROUGE score[5] and BERT score between the 100-shot and fine-tuned model, the authors conclude the 100-shot model to be the best-performing one (as the ΔEI metric is more relevant for this task).

Table 2 Results of the proposed approach; ↑ higher is better, ↓ lower is better

	R-1↑	R-2↑	R-L↑	BERTScore↑	ΔEI-Avg↓	ΔEI-F1↓
0-shot	0.176	0.022	0.111	0.810	0.528	0.387
10-shot	0.190	0.033	0.135	0.830	0.540	0.358
100-shot	0.213	0.038	0.140	0.831	**0.435**	**0.285**
Fine-tune	**0.227**	**0.048**	**0.169**	**0.849**	0.614	0.310

7.2 Ablation Study

Since this paper leverages two components in its proposed approach, namely the pre-training strategy introduced in the PRIMERA paper (Xiao et al., 2022) and the structured representation of data introduced in (DeYoung et al., 2021, p. 2), a simple ablation study was performed wherein the proposed approached (IDEO-PRIMERA) was compared with 1. the model trained without graph encodings (PRIMERA) and 2. the model trained without graph encodings as well as the PRIMERA pre-training strategy (LED-4k). All three models used only 8 studies per review as part of their inputs and have a maximum input length of 4096 tokens to establish a fair comparison.

The results (as presented in Table 3) show that using PRIMERA's pre-training strategy enabled the model to generate better summaries in zero-shot and few-shot settings (while not directly obvious from the metrics, a look into the actual generated summaries revealed that the quality of the proposed method's summaries was significantly better compared to the others. This highlights a drawback as explained in Section 8). Moreover, the addition of graph encodings enabled the model to focus on the crux of each

[5] The average ROUGE (avgr) score is simply the average of the R-1, R-2 and R-L scores.

study of a given review, which in turn allowed it to generate more detailed and factual summaries. Refer to Table 8 of the Appendix for a R-1, R-2 and R-L score comparison of the ablation study.

Table 3 Results of the ablation study; ↑ higher is better, ↓ lower is better

	IDEO-PRIMERA			PRIMERA			LED-4k		
	avgr↑	ΔEI-Avg↓	ΔEI-F1↓	avgr↑	ΔEI-Avg↓	ΔEI-F1↓	avgr↑	ΔEI-Avg↓	ΔEI-F1↓
0-shot	0.103	0.528	0.387	0.101	0.562	0.358	0.046	0.481	0.358
10-shot	0.119	0.540	0.358	0.109	0.560	0.375	0.098	0.540	0.347
100-shot	0.130	**0.435**	**0.285**	0.128	0.475	0.412	0.124	0.535	0.372
Fine-tune	**0.148**	0.614	0.310	0.139	0.619	0.307	0.143	0.595	0.331

7.3 MSLR Shared Task

This section compares the proposed approach with the submitted systems and baselines of the MSLR shared task. The baselines (*) for this task for the MS2 dataset include a Longformer Encoder Decoder (LED) model and BART model trained on the dataset. These models used the background of the review in addition to the study abstracts as part of the input. (*John Giorgi et al. 2022*, n.d.) fine-tuned a LED-16k model following a similar protocol to PRIMERA. (Tangsali et al., 2022) experimented with fine-tuning BART-large, DistillBART and T5-base models on both the MS2 and Cochrane datasets. (Yu, 2022) submitted a Long-T5 model that had been pre-trained on the PubMed dataset.

Table 4 Comparative results of the MSLR shared task; ↑ higher is better, ↓ lower is better

	R-1↑	R-2↑	R-L↑	BERTScore ↑	ΔEI-Avg↓	ΔEI-F1↓
(*John Giorgi et al. 2022*, n.d.) LED-base-16k	**0.275**	**0.092**	**0.206**	**0.869**	0.487	0.424
(Tangsali et al., 2022) PuneICT	0.206	0.035	0.144	0.848	0.532	0.356
(Yu, 2022) LongT5-Pubmed	0.120	0.013	0.096	0.828	0.528	0.343
(DeYoung et al., 2021) Longformer-MS2 *	0.264	0.080	0.196	0.867	0.462	0.412
(DeYoung et al., 2021, p. 2) BART-MS2 *	0.263	0.077	0.195	0.864	0.451	0.414
Proposed method	0.213	0.038	0.140	0.831	**0.435**	**0.285**

Table 4 shows a comparative view of the ROUGE, BERT and EI metrics of all the submitted systems with the proposed approach. It can be observed that while the LED-base-16k and the baselines have much higher ROUGE scores, the proposed system outperforms the existing methods in terms of the ΔEI metrics.

Specifically, the proposed approach shows a 16.90% improvement in the ΔEI-F1 score compared to the next best model in this metric, LongT5-Pubmed. This can be attributed to the fact that by augmenting the input with IDEO tuples, the model was trained to capture the evidence direction D for each intervention I and outcome O using a supporting statement E while generating summaries.

8. Limitations and Future Work

As mentioned in Section 3.1, each study and review abstract are tagged with PICO (Population, Intervention, Comparator and Outcome) elements, as it is known to be a common way of representing clinical knowledge. Building on this, each Intervention and Outcome is also associated with an Evidence and Inference statement. However, all of these steps are done using different classification models as described in (DeYoung et al., 2021, p. 2). These models introduce misclassification errors that can compound at each step of the pipeline (PICO tagging, taking the product of Is and Os at the document level, and predicting EI direction) and influence the performance of the proposed method. A visible example of this can be seen in Table 1 of Section 4.1, where NEC is tagged as an Outcome in the third row but not in the second row. Improvements in the above-mentioned pipeline could greatly improve the model performance.

Another major limitation is the lack of evaluation metrics that can better capture the quality of the generated summary in the domain of Multi-document Summarization (MDS). This has been highlighted in various other studies as well (Ma et al., 2022). Specifically, while ROUGE scores are essential indicators in many NLP tasks, they cannot measure the semantic similarity between generated and reference summaries, which is critical in this domain. On the other hand, while BERTScore does capture semantic meaning to an extent, it also does not fully represent the manner in which cross-document relations are captured in a summary. The evidence inference metrics introduced by (DeYoung et al., 2021, p. 2) are probably currently best suited for the task at hand but they too come with their own challenges. Particularly, these scores are hard to interpret, for e.g., we cannot say how much worse a model with a 0.6 ΔEI-F1 score is compared to one with a 0.4 score. More research is required to develop metrics that can better represent the quality of the generated summaries.

The proposed method in this paper leverages a pre-trained PRIMERA model, whose architecture limits the input size to 4096 tokens. Thus, as described in Section 4.2, each review only includes 8 studies so as to include most of every study's abstract and its corresponding IDEO representation in the input. Theoretically, a PRIMERA model with 16k input length combined with the proposed IDEO attention would be able to include up to 25 studies per review in ins input without loss of information and would thus be able to yield much better results. However, this is limited by the lack of availability of GPU resources.

9. Conclusion

In this paper, the authors tackled the problem of multi-document summarization of medical literature reviews by leveraging a structured form of medical data representation introduced by (DeYoung et al., 2021, p. 2) as well as the pre-training strategy used in the PRIMERA architecture (Xiao et al., 2022). The proposed model was experimented in zero-shot, few-shot and fine-tuned settings and was also compared with other state-of-the-art architectures in the same environments. The proposed method and its comparisons were evaluated using various metrics such as the ROUGE scores, the BERTScore as well the

Evidence-Inference (ΔEI) metrics introduced in (DeYoung et al., 2021, p. 2). Results of the experiments showed that the PRIMERA model, when augmented with the proposed graph attention, outperforms the existing benchmarks in terms of the ΔEI metrics. It also fares well in comparison in both zero-shot and few-shot settings. This paper also highlighted the limitations of the current work in terms of the quality of the annotated dataset as well as that of metrics that can better capture the factuality of the generated summary and raises need for future work in these areas.

References

Amplayo, R. K., & Lapata, M. (2021). Informative and Controllable Opinion Summarization. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2662–2672. https://doi.org/10.18653/v1/2021.eacl-main.229

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The Long-Document Transformer* (arXiv:2004.05150). arXiv. https://doi.org/10.48550/arXiv.2004.05150

Bražinskas, A., Lapata, M., & Titov, I. (2020a). Unsupervised Opinion Summarization as Copycat-Review Generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5151–5169. https://doi.org/10.18653/v1/2020.acl-main.461

Bražinskas, A., Lapata, M., & Titov, I. (2020b). *Few-Shot Learning for Opinion Summarization* (arXiv:2004.14884). arXiv. https://doi.org/10.48550/arXiv.2004.14884

Caciularu, A., Cohan, A., Beltagy, I., Peters, M., Cattan, A., & Dagan, I. (2021). CDLM: Cross-Document Language Modeling. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2648–2662. https://doi.org/10.18653/v1/2021.findings-emnlp.225

Chu, E., & Liu, P. (2019). MeanSum: A Neural Model for Unsupervised Multi-Document Abstractive Summarization. *Proceedings of the 36th International Conference on Machine Learning*, 1223–1232. https://proceedings.mlr.press/v97/chu19b.html

Coavoux, M., Elsahar, H., & Gallé, M. (2019). Unsupervised Aspect-Based Multi-Document Abstractive Summarization. *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 42–47. https://doi.org/10.18653/v1/D19-5405

DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., & Wang, L. (2021). MS^2: Multi-Document Summarization of Medical Studies. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7494–7513. https://doi.org/10.18653/v1/2021.emnlp-main.594

DeYoung, J., Lehman, E., Nye, B., Marshall, I., & Wallace, B. C. (2020). Evidence Inference 2.0: More Data, Better Models. *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 123–132. https://doi.org/10.18653/v1/2020.bionlp-1.13

Fabbri, A., Li, I., She, T., Li, S., & Radev, D. (2019). Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1074–1084. https://doi.org/10.18653/v1/P19-1102

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional Sequence to Sequence Learning. *Proceedings of the 34th International Conference on Machine Learning*, 1243–1252. https://proceedings.mlr.press/v70/gehring17a.html

Goodwin, T., Savery, M., & Demner-Fushman, D. (2020). Flight of the PEGASUS? Comparing Transformers on Few-shot and Zero-shot Multi-document Abstractive Summarization. *Proceedings of the 28th International Conference on Computational Linguistics*, 5640–5646. https://doi.org/10.18653/v1/2020.coling-main.494

Huang, X., Lin, J., & Demner-Fushman, D. (2006). Evaluation of PICO as a Knowledge Representation for Clinical Questions. *AMIA Annual Symposium Proceedings*, *2006*, 359–363.

Jin, H., Wang, T., & Wan, X. (2020a). SemSUM: Semantic Dependency Guided Neural Abstractive Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), Article 05. https://doi.org/10.1609/aaai.v34i05.6312

Jin, H., Wang, T., & Wan, X. (2020b). Multi-Granularity Interaction Network for Extractive and Abstractive Multi-Document Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6244–6254. https://doi.org/10.18653/v1/2020.acl-main.556

John Giorgi et al. 2022. (n.d.). Retrieved December 18, 2022, from https://leaderboard.allenai.org/mslr-ms2/submission/ccfknkbml1mljnftf7d0

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, *96*(3), 118–121.

Kipf, T. N., & Welling, M. (2017). *Semi-Supervised Classification with Graph Convolutional Networks* (arXiv:1609.02907). arXiv. https://doi.org/10.48550/arXiv.1609.02907

Lebanoff, L., Song, K., & Liu, F. (2018). Adapting the Neural Encoder-Decoder Framework from Single to Multi-Document Summarization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4131–4141. https://doi.org/10.18653/v1/D18-1446

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension* (arXiv:1910.13461). arXiv. https://doi.org/10.48550/arXiv.1910.13461

Li, P., Lam, W., Bing, L., Guo, W., & Li, H. (2017). Cascaded Attention based Unsupervised Information Distillation for Compressive Summarization. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2081–2090. https://doi.org/10.18653/v1/D17-1221

Li, W., Xiao, X., Liu, J., Wu, H., Wang, H., & Du, J. (2020). Leveraging Graph to Improve Abstractive Multi-Document Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6232–6243. https://doi.org/10.18653/v1/2020.acl-main.555

Lin, J., Sun, X., Ma, S., & Su, Q. (2018). *Global Encoding for Abstractive Summarization* (arXiv:1805.03989). arXiv. https://doi.org/10.48550/arXiv.1805.03989

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). *Generating Wikipedia by Summarizing Long Sequences* (arXiv:1801.10198). arXiv. https://doi.org/10.48550/arXiv.1801.10198

Liu, Y., & Lapata, M. (2019). Hierarchical Transformers for Multi-Document Summarization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5070–5081. https://doi.org/10.18653/v1/P19-1500

Liu, Y., Luo, Z., & Zhu, K. (2018). Controlling Length in Abstractive Summarization Using a Convolutional Neural Network. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4110–4119. https://doi.org/10.18653/v1/D18-1444

Ma, C., Zhang, W. E., Guo, M., Wang, H., & Sheng, Q. Z. (2022). Multi-document Summarization via Deep Learning Techniques: A Survey. *ACM Computing Surveys*. https://doi.org/10.1145/3529754

Michelson, M., & Reuter, K. (2019). The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary Clinical Trials Communications*, *16*, 100443. https://doi.org/10.1016/j.conctc.2019.100443

Nenkova, A., & Passonneau, R. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 145–152. https://aclanthology.org/N04-1019

Nye, B. E., DeYoung, J., Lehman, E., Nenkova, A., Marshall, I. J., & Wallace, B. C. (2020). *Understanding Clinical Trial Reports: Extracting Medical Entities and Their Relations*. https://doi.org/10.48550/arXiv.2010.03550

Nye, B., Li, J. J., Patel, R., Yang, Y., Marshall, I., Nenkova, A., & Wallace, B. (2018). A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 197–207. https://doi.org/10.18653/v1/P18-1019

Otmakhova, Y., Truong, T. H., Baldwin, T., Cohn, T., Verspoor, K., & Lau, J. H. (2022). LED down the rabbit hole: Exploring the potential of global attention for biomedical multi-document summarisation. *Proceedings of the Third Workshop on Scholarly Document Processing*, 181–187. https://aclanthology.org/2022.sdp-1.21

Pasunuru, R., Liu, M., Bansal, M., Ravi, S., & Dreyer, M. (2021). Efficiently Summarizing Text and Graph Encodings of Multi-Document Clusters. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4768–4779. https://doi.org/10.18653/v1/2021.naacl-main.380

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). *Language Models are Unsupervised Multitask Learners*. 24.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (n.d.). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 67.

See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1073–1083. https://doi.org/10.18653/v1/P17-1099

Song, S., Huang, H., & Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. *Multimedia Tools and Applications*, *78*(1), 857–875. https://doi.org/10.1007/s11042-018-5749-3

Tangsali, R., Vyawahare, A. J., Mandke, A. V., Litake, O. R., & Kadam, D. D. (2022). Abstractive Approaches To Multidocument Summarization Of Medical Literature Reviews. *Proceedings of the Third Workshop on Scholarly Document Processing*, 199–203. https://aclanthology.org/2022.sdp-1.24

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph Attention Networks* (arXiv:1710.10903). arXiv. https://doi.org/10.48550/arXiv.1710.10903

Wallace, B. C., Saha, S., Soboczenski, F., & Marshall, I. J. (2021). Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization. *AMIA Summits on Translational Science Proceedings*, *2021*, 605–614.

Wang, L., & Ling, W. (2016). Neural Network-Based Abstract Generation for Opinions and Arguments. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 47–57. https://doi.org/10.18653/v1/N16-1007

Xiao, W., Beltagy, I., Carenini, G., & Cohan, A. (2022). *PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization* (arXiv:2110.08499). arXiv. https://doi.org/10.48550/arXiv.2110.08499

Xu, H., Wang, Y., Han, K., Ma, B., Chen, J., & Li, X. (2020). Selective Attention Encoders by Syntactic Graph Convolutional Networks for Document Summarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8219–8223. https://doi.org/10.1109/ICASSP40776.2020.9054187

Yu, B. (2022). Evaluating Pre-Trained Language Models on Multi-Document Summarization for Literature Reviews. *Proceedings of the Third Workshop on Scholarly Document Processing*, 188–192. https://aclanthology.org/2022.sdp-1.22

Yuan, C., Bao, Z., Sanderson, M., & Tang, Y. (2020). Incorporating word attention with convolutional neural networks for abstractive summarization. *World Wide Web*, *23*(1), 267–287. https://doi.org/10.1007/s11280-019-00709-6

Yuan, Q., Ni, P., Liu, J., Tong, X., Lu, H., Li, G., & Guan, S. (2021). An Encoder-decoder Architecture with Graph Convolutional Networks for Abstractive Summarization. *2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI)*, 91–97. https://doi.org/10.1109/BDAI52447.2021.9515256

Zhang, J., Tan, J., & Wan, X. (2018). Adapting Neural Single-Document Summarization Model for Abstractive Multi-Document Summarization: A Pilot Study. *Proceedings of the 11th International Conference on Natural Language Generation*, 381–390. https://doi.org/10.18653/v1/W18-6545

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *Proceedings of the 37th International Conference on Machine Learning*, 11328–11339. https://proceedings.mlr.press/v119/zhang20ae.html

Appendix

Table 5 Comparison of Deep Learning methods for Abstractive Summarization

Methods	Works	# Documents		Domain	Pros & Cons
		SDS	MDS		
RNN	(Amplayo & Lapata, 2021)		✓	Reviews	*Pros*: Can capture sequential relationships in sentences *Cons*: Vanishing gradients and information bottlenecks for long sequences
	(Bražinskas et al., 2020a)		✓	Reviews	
	(Chu & Liu, 2019)		✓	Reviews	
	(Coavoux et al., 2019)		✓	Reviews	
	(P. Li et al., 2017)		✓	News	
	(Wang & Ling, 2016)		✓	Reviews	
	(Zhang et al., 2018)		✓	News	
CNN	(Lin et al., 2018)	✓		News	*Pros*: Allows for parallel computing *Cons*: Cannot capture long-term dependencies
	(Y. Liu et al., 2018)	✓		News	
	(Song et al., 2019)	✓		News	
	(C. Yuan et al., 2020)	✓		News	
GNN	(Jin et al., 2020a)	✓		News	*Pros*: model semantic and syntactic relationships between entities in natural language *Cons*: Inefficient for large graphs
	(Xu et al., 2020)	✓		News	
	(Q. Yuan et al., 2021)	✓		News	
PG	(Fabbri et al., 2019)		✓	News	*Pros*: Low redundancy *Cons*: Hard to train
	(Lebanoff et al., 2018)		✓	News	
Transformer	(Bražinskas et al., 2020b)		✓	Reviews	*Pros:* Can handle long sequences efficiently; capture in-document and cross-document relations; allow for parallelization *Cons:* Memory-intensive and time-consuming to train
	(Jin et al., 2020b)		✓	News	
	(W. Li et al., 2020)		✓	Wikipedia	
	(P. J. Liu et al., 2018)		✓	Wikipedia	
	(Y. Liu & Lapata, 2019)		✓	Wikipedia	
	(Pasunuru et al., 2021)		✓	News	
	(Xiao et al., 2022)		✓	News, Science, Wikipedia	
	(Zhang et al., 2020)	✓	✓		

Table 6 Statistics of studies per review and IDEOs per study

Statistic	Train		Validation		Test	
	Studies per review	IDEOs per study per review	Studies per review	IDEOs per study per review	Studies per review	IDEOs per study per review
Mean	26.31	3.43	26.28	3.44	28.49	3.3
Max	390	60	305	70	292	60
Min	1	1	2	1	1	1
SD	28.12	3.69	26.61	4.64	29.15	4.15
Q1	11	1	11	1	12	1
Q2	18	2	19	2	20	2
Q3	31	4	30	4	34	4

Table 7 Statistics of abstract and IDEO encoding lengths (per review)

Statistic	Train		Validation		Test	
	Abstract encoding length	IDEO encoding length	Abstract encoding length	IDEO encoding length	Abstract encoding length	IDEO encoding length
Mean	376.26	210.18	376.05	216.83	380.43	205.21
Max	1173.5	3880	880	5128.40	1316.4	3373.88
Min	108	32	187.53	39.75	157.5	37.71
SD	81.42	226.96	68.69	315.01	81.94	254.37
Q1	324.50	75	331.35	73.12	334.53	73.05
Q2	372.72	133.23	376.25	129.14	373.1	129.44
Q3	420.32	242.41	411.89	236.29	415.96	225.22

Table 8 ROUGE score comparison of the ablation study

	IDEO-PRIMERA			PRIMERA			LED-4k		
	R-1 ↑	R-2 ↑	R-L ↑	R-1 ↑	R-2 ↑	R-L ↑	R-1 ↑	R-2 ↑	R-L ↑
0-shot	0.176	0.022	0.111	0.174	0.019	0.110	0.073	0.008	0.059
10-shot	0.190	0.033	0.135	0.169	0.028	0.130	0.158	0.019	0.117
100-shot	0.213	0.038	0.140	0.204	0.034	0.146	0.197	0.032	0.144
Fine-tune	**0.227**	**0.048**	**0.169**	0.217	0.039	0.162	0.226	0.042	0.163