

# Analysis of GraphSum’s Attention Weights to Improve the Explainability of Multi-Document Summarization

M. Lautaro Hickmann  
Ulm University  
lautaro.hickmann@uni-ulm.de

Fabian Wurzberger  
Ulm University  
fabian.wurzberger@uni-ulm.de

Megi Hoxhalli  
Ulm University  
megi.hoxhalli@uni-ulm.de

Arne Lochner  
Ulm University  
arne.lochner@uni-ulm.de

Ansgar Scherp\*  
Ulm University  
ansgar.scherp@uni-ulm.de

## ABSTRACT

Modern multi-document summarization (MDS) methods are based on transformer architectures. They generate state of the art summaries, but lack explainability. We focus on graph-based transformer models for MDS as they gained recent popularity. We aim to improve the explainability of the graph-based MDS by analyzing their attention weights. In a graph-based MDS such as GraphSum, vertices represent the textual units, while the edges form some similarity graph over the units. We compare GraphSum’s performance utilizing different textual units, i. e., sentences versus paragraphs, on two news benchmark datasets, namely WikiSum and MultiNews. Our experiments show that paragraph-level representations provide the best summarization performance. Thus, we subsequently focus on analyzing the paragraph-level attention weights of GraphSum’s multi-heads and decoding layers in order to improve the explainability of a transformer-based MDS model. As a reference metric, we calculate the ROUGE scores between the input paragraphs and each sentence in the generated summary, which indicate source origin information via text similarity. We observe a high correlation between the attention weights and this reference metric, especially on the later decoding layers of the transformer architecture. Finally, we investigate if the generated summaries follow a pattern of positional bias by extracting which paragraph provided the most information for each generated summary. Our results show that there is a high correlation between the position in the summary and the source origin.

## KEYWORDS

similarity graph, abstractive MDS, textual units

## 1 INTRODUCTION

Multi-document Summarization (MDS) refers to the task of providing a concise representation of multiple documents with overlapping textual content [14]. Besides transformer-based models for summarization [16, 17], approaches on graph neural networks and specifically using knowledge graphs have gained popularity [6, 12, 20, 21] while more recent models combine graph neural networks with a transformer architecture [6, 12].

Generally, we distinguish extractive and abstractive MDS. In *extractive MDS*, a model is trained to select the most relevant input sentences or paragraphs in order to create an optimal summary with

regard to salience and coherence [21]. In comparison, *abstractive MDS* is able to generate new texts which are not present in the input documents, which improve the summary quality [6]. For extractive MDS, a graph-based representation can help to detect salient vertices, so that they are included in the generated summary to cover all important information. In abstractive MDS, salient vertices can guide the summarization process via attention mechanisms of the transformer architecture. To this end, the text from multiple documents is represented as independent, so-called *textual units*. The graph’s vertices represent the textual units such as paragraphs and sentences, while the edges model a semantic relationship between the units like a text-based cosine similarity. Different textual units have been used for graph-based text summarization in the past, including discourse relations [6, 20], sentences [2, 21], and paragraphs [12].

Recently, Li et al. [12] proposed GraphSum, a state of the art graph-based MDS, which utilizes a graph structure based on paragraphs as textual units to improve the transformer architecture and guide the summary generation process [12]. Paragraphs are assumed to divide text into contiguous topics. Leveraging inter-paragraph relations can provide the model additional information for detecting contextual relations between topics. The GraphSum model showed substantial improvements compared to strong MDS baselines such as LEAD-3 and LexRank [4].

The authors of GraphSum [12] did consider paragraph-level textual units as vertices for their graph structure. Also Huang et al. [6] and others used paragraph-level representations for the graph-based MDS task. Other works for MDS considered sentence-level textual units for the summarization tasks such as [2, 19, 21]. However, to the best of our knowledge, there is no graph-based MDS model that has used representation of textual units on both levels, i. e., paragraph-level and sentence-level, on the same dataset. Thus, in a first step, we investigate if a finer-grained text representation with textual units on sentence level improves the results. We focus on using GraphSum as state of the art model to contrast the summarization results of a representation on paragraph level vs sentence level. Our rationale is that with sentences as textual units, the graph structure represents inter-sentence relations, which may provide more detailed information within topics and thus may improve the results.

Based on the first step, we are interested whether a graph-based MDS model extracts information from textual units based on their position in the content. It would mean that the summarization model is capable of picking up this bias from the input data. Such a

\* Author is contributing Conceptualization, Writing - Review & Editing, and Supervision. Statement is based on the Contributor Roles Taxonomy, see: <http://credit.niso.org/>

*source origin analysis* is particularly interesting for datasets where the gold standard is biased towards selecting information from the first sentences. In example, for MDS tasks within the news domain, Kedzie et al. [8] and Jung et al. [7] suggest that news summarizations follows a *positional bias* pattern. In the news domain, the information from the first sentences of the input documents is the most relevant. This is the main reason why the simple, *extractive* summarization model LEAD-3, where the first 3 sentences are selected to generate a summary, performs well as an model. However, the LEAD-3 model is biased towards the news domain. It exploits the specific knowledge about how journalists write their articles. In abstractive MDS, such as GraphSum, new sentences are generated and therefore no explainable source origin information can be directly extracted from the generated summaries. However, abstractive transformer-based MDS models often use positional encodings [18] to make use of the sequential structure of the data. Thus, in the second step we analyze the attention weights in the GraphSum model to improve the understanding of abstractive MDS.

In summary, we investigate the following two research questions:

- (1) We construct similarity graphs on sentence-level and paragraph-level in order to guide a graph-transformer-based MDS model [15]. We investigate which of the aforementioned textual units are most beneficial as vertices in a similarity graph to capture context and concepts of the input documents and therefore guide the MDS process most effectively.
- (2) We analyze attention mechanisms in an abstractive graph-transformer-based MDS model [15] in order to detect which specific textual units provide the most valuable information for a generated summary. We use the two benchmark datasets MultiNews and WikiSum. This shall shed light onto the decisions made by the MDS regarding the positions of the text units within the larger documents, i. e., make more explainable from which part of the input the resulting summaries are generated.

Thus, overall the contribution of this work is to better understand how similarity graphs guide the summarization process. Regarding the first step, our experiments show that sentence-level representations of textual units in GraphSum are outperformed by paragraph-level representations for MDS on the MultiNews dataset. We focus on the news domain, and particularly the benchmark dataset MultiNews, in preparation to the second step on the analysis of the attention weights as motivated above. Regarding the results of the second step, we analyze the attention weights of GraphSum on two benchmark datasets, namely MultiNews and WikiSum. Our results indicate that the attention weights of later decoding layers of the transformer architecture can provide source origin information. This suggests that attention weights can be used in further work to analyze if positional information (e. g., positional encoding of textual units) in abstractive transformer-based MDS leads to the side effect, that trained models learn a positional bias as mentioned previously.

The remainder of the article is organized as follows. Below we discuss related work on MDS, with specific focus on methods that leverage graph structures. In Section 3, we introduce the methods and data pre-processing two answer the two research questions.

In Section 4, we describe the experimental apparatus. Section 5 presents the experimental results. We discuss the results in Section 6, before we conclude.

## 2 RELATED WORK

Multi-document Summarization (MDS) can be viewed as a sequence-to-sequence task [12], where a set of multiple input documents is transformed into a single text summary. Most abstractive MDS models were based on an auto-regressive encoder-decoder architecture, where the encoder maps an input sequence to a hidden representation and the decoder transforms the hidden representation into another sequence in a token-by-token manner [18]. More elaborate approaches further included an attention mechanism in the encoder-decoder architecture, which improves the models when using long sequences [1, 18].

Recently graph structures were used to improve the performance of extractive and abstractive MDS models by guiding the summarization process implicitly and explicitly. Similarity and discourse graphs have been used for MDS [12]. Especially similarity graphs, which are able to capture lexical or topic relations between textual units, improve the quality of the generated summaries [12]. Below, we review the state of the art in graph-based extractive and graph-based abstractive MDS approaches.

### 2.1 Extractive Graph-based Summarization

An early work from 2013 by Christensen et al. [2] proposed to leverage graph-structures for extractive text summarization in order to improve the coherence of the generated summary. They constructed a directed discourse graph with input sentences as vertices, where the edges between two vertices represent constraints for sentence ordering to achieve a coherent summary. The process of retrieving edge weights for this discourse graph was based on heuristic metrics such as deverbal noun references, event-entity continuation, discourse markers, inferred edges, and co-referent mentions. These metrics were used to construct the Approximate Discourse Graph (ADG). All possible candidate summaries for the input documents were evaluated based on a joint objective function, where coherence and salience were a soft constraint. The ADG enabled the MDS model to estimate the coherence score of a candidate summary by analyzing the edge weights of successive summary sentences. A linear regression model was trained based on ROGUE scores to predict the salience score of a candidate summary. The candidate summary, which maximized this function was returned as the generated summary.

As follow up work, Yasunaga et al. [21] improved the creation process of the ADG in 2017. While Christensen et al. [2] used the graph structure only implicitly for constraint purposes, Yasunaga et al. [21] used the graph structure explicitly for a Graph Convolution Network (GCN) [9]. Yasunaga et al. noticed that the absolute value of the edge weights in Christensen et al.’s ADG lacked diversity, because they were only incremented by discrete values and many edge weights were equal to one. Thus, the edge weights did not provide additional information for the summarization beyond the simple knowledge about the presence of a relation between two sentences. To overcome this issue, Yasunaga et al. first transformed the edge weights for each sentence via sentence personalization and

afterwards normalized over all incoming edges. This resulted in the Personalized Discourse Graph (PDG), which reflected macro-level sentence features. They were helpful to estimate salience information. In the experiments, PDG performed better than ADG with regards to ROUGE scores, learning curves, and computation costs. Additionally, the node degree in PDG has had a higher correlation to the salience score estimated by the linear regression than its counterpart ADG.

More recently, Wang et al. [19] presented a new architecture named HETERSUMGRAPH for extractive summarization based upon a heterogeneous graph-based neural network, built by nodes of varying granularity. These nodes were then used to enrich the cross sentence relations, making it flexible to use in single-document and multi-document summarization. The model consists of three parts: graph initializer, heterogeneous graph layer, and sentence selector. To create the sentence nodes, they used a concatenation of a CNN and BiLSTM, and for the relations between them, they made use of TF-IDF. For the heterogeneous graph layer, HETERSUMGRAPH used Graph Attention Network (GAT), and simple feed-forward neural network in an iterative process. Finally, in the sentence selecting process, trigram blocking was used to score and discard redundant sentences.

## 2.2 Abstractive Graph-based Summarization

While Christensen et al. [2] and Yasanaga et al. [21] already leveraged discourse graphs in an implicit and explicit manner, Li et al. [12] extended this idea by using similarity graphs based on TF-IDF as well as topic graphs in a transformer-based architecture [18]. The graph structure of the input data was used to guide the attention mechanism within the encoder and decoder of the transformer-like architecture. This allowed the adjacency-matrix to explicitly influence the encoder and decoder part of the summarizer. The resulting system was called *GraphSum*. The graph encoding layers applied graph-informed self-attention onto the input data, where the attention weights are shifted by the graph-information. The decoder within the GraphSum model contained a Global Graph Attention mechanism. This mechanism introduced a global context vector, which is a weighted sum of the central textual unit vectors. In addition, the decoder contained a local normalized attention, in which each input token is attended separately. The inference process of GraphSum was extended by a beam search in order to improve the quality of generated summaries. For evaluation of GraphSum, Li et al. [12] used the WikiSum-Dataset as well as the MultiNews-Dataset. For both datasets, paragraphs acted as the central textual units for the graph-structure. Sentence-level representations have not been considered.

Another approach is ASGARD by Huang et al. [6]. ASGARD used GATs to capture global context based on Knowledge Graphs (KGs), which are constructed by an openIE model. The authors argued that topics are naturally split by paragraphs, therefore they also employed their GAT encoding on paragraph level by sampling sub-graphs from the KG. Furthermore, they also followed the commonly used encoder-decoder architecture. Additionally, they used graph attention and token level attention in the decoder part. Besides next token prediction, Huang et al. introduced the additional task of node salience labeling, where the model is trained to predict a

salience score for each node. Those scores were based on the occurrences of nodes in the gold standard and were used to filter nodes in the graph encoder. In addition, to optimize the cross entropy loss for the next token prediction, Huang et al. further assumed that informative summaries yield better accuracy on question answering tasks. Accordingly, they defined a reinforcement objective based on training a question answering model on the generated summaries.

## 2.3 Summary

In contrast to existing works, we investigate the influence of different textual units i. e., using sentence-level vs. paragraph-level textual units, for training the same abstractive graph-based MDS model. In addition, we investigate the possibility to utilize attention weights of a transformer-based decoder in order to extract source origin information of a generated summary.

Our analysis is based on the GraphSum [12] model, which delivers state-of-the-art results for graph-transformer-based MDS architectures, enabling a source origin analysis via attention weights. GraphSum is extended by pre-processing the input data to represent sentence-level and paragraph-level textual units. In addition, we extract the generated global attention weights of a trained GraphSum model in order to analyze the possibility of leveraging attention weights to improve the explainability of abstractive MDS.

## 3 EXTENSION OF THE GRAPHSUM MDS

We base on the state of the art abstractive graph-based transformer GraphSum from Li et al. [12].<sup>1</sup> GraphSum uses a graph-based representation of the input paragraphs for the encoder and decoder part of the transformer architecture. Thus, GraphSum is optimally suited to be extended to incorporate a sentence-level representation, to answer the first research questions. Additionally, the extraction of attention weights enables the possibility to analyze source origin relations, to answer the second research question.

Below, we first describe the modification of GraphSum to support both paragraph-level and sentence-level text unit representation. Subsequently, we explain how we extract the attention weights for obtaining the positional information to make the summaries more transparent.

### 3.1 Modifying GraphSum for Supporting Sentence-level Text Representations

The input data of the GraphSum model is tokenized using a language model, namely *sentencepiece* [10]. Then truncating and padding are applied to each input textual unit, as done by GraphSum Li et al. [12]. This results in each input set of multi-documents containing a fixed number of textual units and tokens within each unit. The positional information of each unit is preserved by utilizing positional encoding. This basic pre-processing transforms the input data into a sequence of tokens. This allows us to train GraphSum with minor extensions on both sentence-level and paragraph-level document representations.

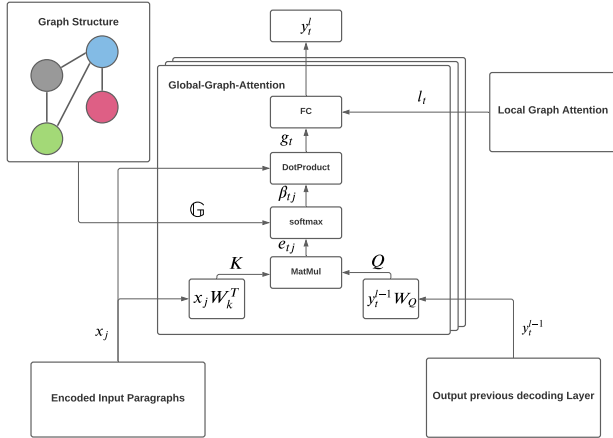
In order to compare the results of the paragraph-level model and sentence-level model, we need to make sure that the sentence-level models receives the same number of tokens compared to

<sup>1</sup>We use the code provided by the authors, which is available here: <https://github.com/PaddlePaddle/Research/tree/master/NLP/ACL2020-GraphSum>

he paragraph-level model. While the total input to the model on sentence-level has the same length, the actual sequence of tokens can slightly differ between the two levels of textual units due to difference in the truncation and padding procedure. The already pre-processed paragraph-level dataset [12] is split into sentence with the help of a tokenizer, namely *sentencepiece* [10]. This will result in an increase of textual units, because each paragraph must contain at least one sentence. The newly obtained textual units differ with regard to the number of tokens. Therefore, the data is truncated and padded to a fixed number of input sentences and a fixed number of tokens for each sentence. The size of the data immensely increases when applying padding on sentence-level text representation (as it is done for the paragraph-level representation), due to the higher number of textual units. The padding is an important factor, as it defines the maximal number of tokens a textual unit can have. Therefore, for sentence-level representation the maximal number of tokens per sentence, i. e., per textual unit is reduced, in order to increase the maximal total number of textual units. Afterwards we build a TF-IDF graph-structure to capture similarities between the newly obtained sentences. The already pre-processed paragraph-level dataset already contains a TF-IDF graph-structure. Therefore, no further pre-processing is necessary for the paragraph-level model.

### 3.2 Extracting the Source Origin via GraphSum’s Attention Weights

GraphSum has in total 8 decoding layers with 8 multi-heads each. A detailed description of the architecture can be found in [12]. Below we focus on explaining the global graph attention as we use this information for the analysis of the source origin information, i. e., understanding the position of the paragraphs that most influence the summaries generated by GraphSum.



**Figure 1: Global graph attention mechanism within a single decoding layer. Multi-Heads are indicated by the stacked boxes (center).**

The global graph attention as described by Li et al. [12] is illustrated in Figure 1. The global context vector  $g_t$  within a decoding

layer  $l$  and the summary token at position  $t$  is obtained by calculating a weighted sum of the attention distribution  $\beta_{tj}$  and the encoded paragraphs’ vectors  $x_j, j \in \{1, \dots, L\}$ , where  $L$  denotes the number of input paragraphs:

$$g_t = \sum_{j=1}^L \beta_{tj} x_j \quad (1)$$

The Attention Weight Distribution (AWD) of decoding layer  $l$  is calculated by predicting the central paragraph  $s_t$  for the attended token  $y_t^{l-1}$  of the previous decoding layer. At each step of the auto-regressive decoder architecture, the central paragraph  $s_t$  represents the paragraph, which is at central position for the next predicted token. The central paragraph  $s_t$  helps to guide the model to attend paragraphs, which are similar to  $x_{s_t}$  according to the graph structure  $G$ . This central position is obtained by transforming the attended token  $y_t^{l-1}$  by a two-layered feed forward network. Note that the central paragraph is solely based on  $y_t^{l-1}$ , which corresponds to the query token of the current attention mechanism.

Afterwards, the attention weights between encoded paragraph  $x_j$  and token vector  $y_t$  are shifted by leveraging graph information between the central paragraph and paragraph  $x_j$ . These shifted attention weights are then transformed into a probability distribution by applying a softmax function resulting in the AWD denoted by  $\beta$ :

$$\beta_{tj} = \text{softmax}(e_{tj} - \frac{(1 - G[s_t][j]^2)}{2\sigma^2}) \quad (2)$$

The scaling factor  $\sigma$  denotes the influence of the graph structure on the attention weights [12]. The unscaled attention weights  $e_{tj}$  between input paragraph  $x_j$  and summary token  $y_t^{l-1}$  are calculated according to Vaswani et al. [18], where  $d_{head}$  denotes the number of multi-heads:

$$e_{tj} = \frac{(y_t^{l-1} W_Q)(x_j W_K)^T}{\sqrt{d_{head}}} \quad (3)$$

After training the GraphSum model as described in [12], we generate summaries for each set of multi-documents from the test dataset and extract the AWD for each generated token, multi-head, decoding layer, and beam. This results in a tensor  $A \in \mathbb{R}^{bs \times sl \times dl \times mh \times L}$  for a single generated summary, where  $bs$  is the number of beams,  $sl$  the length of the generated summary,  $dl$  is the number of decoding layers,  $mh$  the number of multi-heads and  $L$  denoting the number of input paragraphs. Since GraphSum uses beam search for its inference process,  $A$  needs to be transformed by a beam search decoder in order to obtain the correct AWD at each step for the generated summaries.

As the auto-regressive decoder architecture generates a single token with corresponding AWD at each step, this information can only be used to provide source origin information for each token individually. But the source origin information of a single token within a generated summary might not be interesting. Instead it may be desirable to obtain source origin information about a sequence of tokens (e. g., sentences). For this purpose we split the generated summaries into the individual, generated sentences according to a special token which indicates the end of a sentence [11]. This

information is then used to aggregate the token-level information of  $\mathbf{A}$  to sentence-level by applying an aggregation method (e.g., mean or median) resulting in  $\mathbf{A}'$ . Thus, the AWD values of a single paragraph  $x_j$  are an aggregation over all tokens of each generated sentence.

To analyze the possibility of leveraging attention weights for source origin information, we compute the correlation between  $\mathbf{A}'$  and a source origin metric  $\mathbf{R}$ , which is based on the similarity score between the generated summary and the input paragraphs. We use ROUGE scores, as described in Section 4.4.

## 4 EXPERIMENTAL APPARATUS

### 4.1 Datasets

We use two datasets, MultiNews and WikiSum, which are used as benchmarks in evaluating MDS models. MultiNews [5] contains human written summaries from professionals and includes links to the original articles. The articles were taken from newser.com and comprises about 60,000 documents. We use the pre-processed data from Li et al. [12].

WikiSum is a MDS dataset based on references found in Wikipedia articles and Google searches of the Wikipedia articles’ titles [15]. The gold standard summary is the corresponding Wikipedia article itself. WikiSum contains about 2.3 million Wikipedia articles. We will use the ranked version of the WikiSum dataset, which was suggested by Liu et al. [16]. The ranked dataset contains the top-40 paragraphs, which were predicted by a paragraph ranker to be most relevant as described by Liu and Lapata [16].

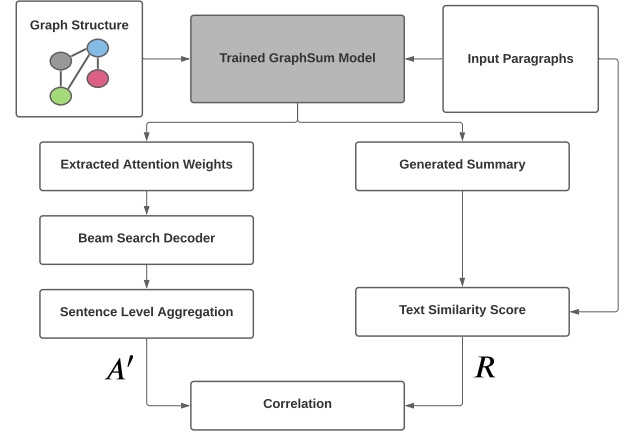
We use the MultiNews dataset for sentences vs. paragraphs comparison. Due to resource restriction, we do not consider the WikiSum dataset for the sentences vs. paragraphs comparison. We use both datasets, i.e., MultiNews and WikiSum, for analyzing the source origin information via attention weights.

### 4.2 Procedure

*Summarization using sentences vs. paragraphs.* Regarding the comparison of sentence-level vs. paragraph-level representation of the input documents, we train our GraphSum models on batch sizes of 3,000. The batch size represents the number of target tokens. Thus, we use the same batch size for both representation to make a fair comparison of the results. Additionally, for the paragraph-level model, we also use a batch size of 4,000, which is not possible anymore for the sentence-level model due to the higher memory requirements. We conduct this comparison on the MultiNews dataset. We split the dataset into 44,927 samples for training, 5,622 for validation, and 5,622 for testing.

*Source origin information via attentions weights.* Regarding the second research question, i.e., analyzing the possible correlation between source origin information and the attention weights, we use the pre-trained paragraph-level GraphSum model provided by Li et al. [12]. In the inference process on the test data, we first extract the multi-head attention weights of the global graph attention  $\mathbf{A}$  within the decoding layers. Afterwards, we apply a beam search decoding and a sentence-level aggregation to achieve  $\mathbf{A}'$ . For the reference metric  $\mathbf{R}$ , we compute a text similarity score between the generated sentences and the input paragraphs by calculating the

ROUGE score. After obtaining  $\mathbf{R}$ , we can calculate the correlation between  $\mathbf{R}$  and  $\mathbf{A}'$ . This procedure is illustrated in Figure 2.



**Figure 2: Processing pipeline to calculate correlation between aggregated AWD  $\mathbf{A}'$  and reference text similarity metric  $\mathbf{R}$ .**

The specific measures taken are described in Section 4.4. Due to resource limitations, we perform our analysis on a reduced test dataset of 500 generated summaries for MultiNews as well as WikiSum. However, in pre-experiments we did not observe a difference when using only a fraction of this number of summaries.

All models are trained on three P100 GPUs with 16GB memory each. The parallel training results in a speed up of the training procedure by increasing the effective batch size. This can also be interpreted as a gradient accumulation over three steps, when combining the three parallel processes. Preliminary results indicate that both models converge within 100 epochs. Therefore, we limit training to 100 epochs with no early stopping.

### 4.3 Hyperparameter Optimization

*Summarization using sentences vs. paragraphs.* As stated in section 3, the input data of the GraphSum model is first truncated and then padded to a fixed number of textual units as well as a fixed number of tokens within each textual unit. Just increasing the amount of textual units when training on sentence-level will result in a noticeable increase in memory consumption. Therefore, it is important to find a good balance between number of textual units and token per textual unit. For paragraph-level representation, GraphSum [12] used 30 paragraphs with a maximum of 60 tokens, resulting in a sequence of 1,800 tokens for each set of multi-documents. For reasons of fair comparison, we chose the parameters for the sentence-level representation such that we have similar complexity. We use 60 sentences and with a maximum of 30 tokens.

Even though the GraphSum model operates on 1,800 tokens for the sentence-level representation as well as for paragraph-level representation, the memory consumption for training on sentence-level is higher. This behaviour arises from a larger number of vertices in the similarity graph, which also needs to fit into the GPU

memory. Therefore, the batchsize when training on sentence-level has to be reduced by 25% in comparison to paragraph-level. Apart from this, we used the same values for the hyperparameters as Li et al. for training the model (e. g., learning rate, beam size, optimizer, among others), which resulted in the best model performance.

*Source origin information via attention weights.* As introduced in Section 3.2, an aggregation metric is applied to transform the AWD containing token-level information to sentence-level. For this purpose we used the mean function.

#### 4.4 Measures

For sentence-level vs. paragraph-level comparison, we use the ROUGE metric to evaluate the quality of the generated summaries against a reference summary. We consider ROUGE-1, which considers overlaps between uni-grams, ROUGE-2, which considers overlaps between bi-grams, and ROUGE-L, which considers the longest common sub sequence [13].

For our source origin analysis, we calculate the ROUGE scores between the input paragraphs and each sentence in the generated summary. Those ROUGE scores calculate the text similarity between the generated sentence and the input paragraphs. To this end, we split each input paragraph into its sentences and separately calculate the ROUGE score for each sentence. Those ROUGE scores are then averaged to obtain a paragraph-level  $\mathbf{R}$  score. This score indicates the source origin information of the generated sentence based on the text similarity to the input paragraphs. After obtaining the source origin metric  $\mathbf{R}$ , we calculate Pearson’s correlation coefficient between the aggregated attention weights  $\mathbf{A}'$  and  $\mathbf{R}$ , for each decoding layer and multi-head to evaluate how effectively attention weights in a transformer-based decoder can indicate source origin information.

## 5 RESULTS

### 5.1 Summaries Using Sentences vs. Paragraphs

Table 1 shows the results for the comparison of sentence-level vs. paragraph-level summaries generated by our MDS models on the test dataset. With a batchsize of 3,000 target tokens, the paragraph-level model outperforms the sentence-level model with regard to the ROUGE score by a small margin. Furthermore, a larger batchsize of 4,000 resulted in higher ROUGE scores for the paragraph-level models. These scores are similar to the best results reported by Li et al., which confirms the validity of our results.

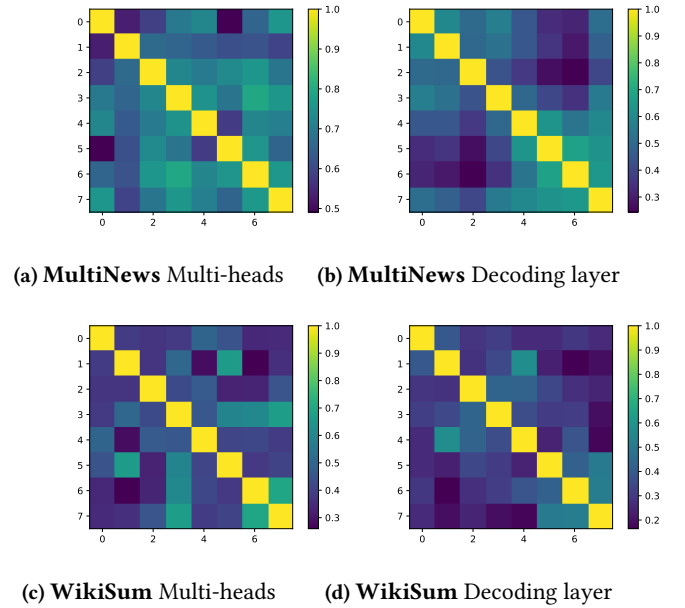
Textual Unit	ROUGE-F (1/2/L)	Batch Size (number of target tokens)
Sentences	43.82/15.85/40.31	3,000
Paragraphs	44.82/16.89/41.17	3,000
Paragraphs	45.06/16.84/41.35	4,000
Li et al. [12]	45.71/17.12/41.99	4,096

**Table 1: Mean test results over 3 runs (ROUGE-F) for different model configurations trained on the MultiNews dataset.**

### 5.2 Source Origin Analysis

The GraphSum model contains 8 decoding layers, while having 8 multi-heads for each attention mechanism. Thus, we analyze if the different multi-heads and decoding-layers attend different input paragraphs for the generation process. Our results show that the different multi-heads have a high correlation between each other with regard to their attention weights. This was especially present in the MultiNews dataset, for which this pattern is shown in Figure 3a. For the MultiNews dataset, the average correlation coefficient is 0.72, while the minimum correlation coefficient between two multi-heads is 0.49. For the WikiSum dataset, shown in Figure 3c, the average coefficient is 0.51, while the minimum correlation is 0.27.

Based on this result, we aggregated the attention weights of different multi-heads within each decoding layer for further analysis with regard to the correlation between the attention weights  $\mathbf{A}'$  and ROUGE-scores  $\mathbf{R}$  (see Section 3.2). The correlation coefficient between different decoding layers varies. For the MultiNews dataset, shown in Figure 3b, the average correlation coefficient is 0.53, while the minimum correlation coefficient is 0.24. For the WikiSum dataset (see Figure 3d) the average correlation is 0.41, while the minimum correlation coefficient is 0.16. As shown in Figure 3b, the correlation between subsequent layers in the MultiNews dataset is the highest. This pattern is not as pronounced in the WikiSum dataset.



**Figure 3: Correlation between the attention weights of different multi-heads (left) and decoding layers (right), evaluated on the MultiNews and WikiSum dataset based on the global attention weights. The attention weights are extracted from a reduced test set.**

The correlation between the aggregated AWD of each decoding layer and the ROUGE-scores  $\mathbf{R}$  are shown in Table 2. These results show that the decoding layers form two clusters. The first



Decoding layer	Correlation( $A'$ , $R$ )	
	MultiNews	WikiSum
1	0.18/0.20/0.19	0.21/0.20/0.21
2	0.16/0.18/0.17	<i>0.02/0.02/0.02</i>
3	<i>0.13/0.15/0.14</i>	0.22/0.23/0.23
4	0.32/0.36/0.34	0.10/0.12/0.11
5	0.46/0.53/0.50	0.15/0.19/0.17
6	<b>0.56/0.69/0.63</b>	0.41/0.52/0.46
7	0.56/0.69/0.63	0.46/0.57/0.52
8	0.47/0.54/0.51	<b>0.49/0.58/0.53</b>

**Table 2: Correlation for the different decoding layers for the MultiNews dataset and WikiSum dataset, based on the global attention weights ( $A'$ ) with regard to text-similarity information obtained by ROUGE-scores ( $R$ ), which contains information about ROUGE-F(1/2/L) information. Highest scores are shown in bold, lowest in italics. The attention weights and ROUGE-scores are extracted from a reduced test set.**

cluster contains the first layers, which average a lower correlation coefficient compared to the later layers, which make up the second cluster.

For each cluster, we consider a single layer for further analysis, namely the layer with the lowest correlation (italics) within the first cluster and the layer with the highest correlation (bold font) within the second cluster. The correlation between  $A'$  with their respective  $R$  is visualized in Figure 4 for the MultiNews dataset and in Figure 5 for the WikiSum dataset. The correlation for further decoding layers can be found in Appendix A.

Overall, we observe a high correlation coefficient between the attention weights and reference metric for the source origin analysis. This holds particularly true for the deeper decoding layers as shown in Table 2. Thus, we investigate if the generated summaries of GraphSum follow the pattern of positional bias. For this purpose, we can use the attention weights to retrieve the input paragraph, which provides the most information for a generated sentence. To analyze if the positional bias is present in the generated summaries, we aggregate this information over all generated summaries. Since the input of the GraphSum [12] model is interpreted as a sequence of tokens, we need information about which paragraphs correspond to which input document. The WikiSum dataset does not contain this information, therefore our analysis of positional bias is limited to the MultiNews dataset.

According to the attention weights, for most generated sentences of the summary, the first paragraphs of each input document are attended the most. This pattern can be seen in the heat-maps shown in Figure 6 for layer 3, with the lowest correlation, and layer 6, with the highest correlation. The heat-maps visualize where the information about which paragraph was attended the most for a generated sentence according to the attention weights. Note that the generated summaries have different numbers of sentences. This results in a bias as there are fewer summaries with a high number of sentences in the test set. This behaviour can especially be observed in Figure 6 (b), where artefacts can be seen for later sentences of

generated summaries with a high number of sentences i. e., in the right columns of the heatmap. The plots for further decoding layers can be found in Appendix B.

## 6 DISCUSSION

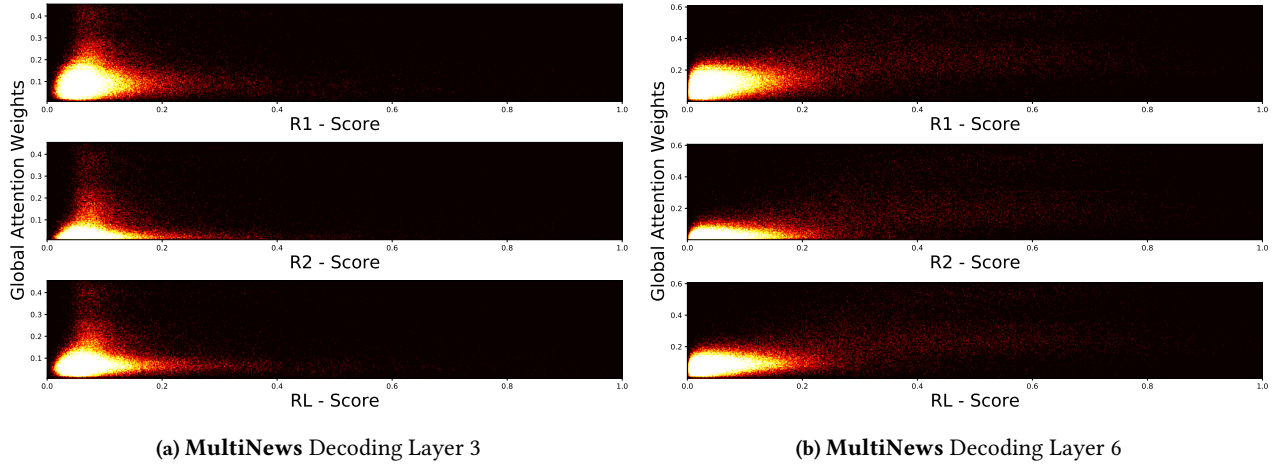
### 6.1 Summaries Using Sentences vs. Paragraphs

Our experiments indicate that paragraph-level models perform better than sentence-level models on the MultiNews dataset. However, the margin is rather small. We have analyzed the MultiNews dataset in more detail. We observe that the paragraphs often contain a single sentence as shown by the long-tail distribution in Figure 7. The y-axis of the right plot is log-scaled in order to visualize the outliers. On average, a paragraph contains only 1.84 sentences. The vast majority of the paragraphs contain less than three sentences. However, there are also outliers in the sense that the paragraphs contain a lot of sentences.

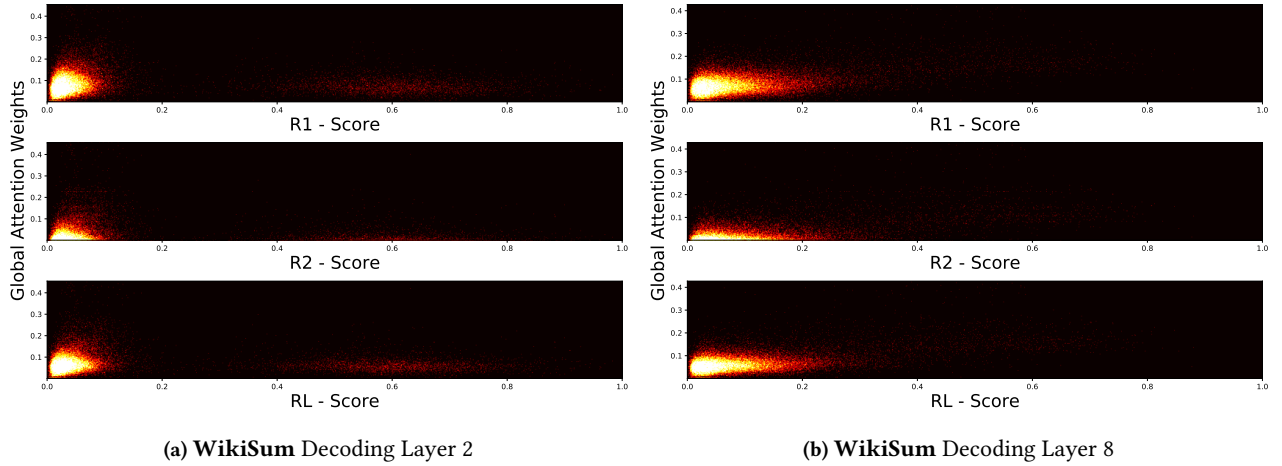
It appears that paragraphs are mainly used as structural aid in news articles to improve readability and not to separate topics, resulting in a low number of sentences per paragraph. This may explain why a sentence-level representation and the paragraph-level representation perform quite similar. We would expect similar observations in other MDS domains where paragraphs are rather short, too.

As the number of sentences in the MultiNews dataset exceeds the number of paragraphs by a factor of 1.84, more textual units have to be considered when working with sentence-level representations to cover the same information. To keep the complexity of the model similar to paragraph-level representations, the number of tokens per textual unit had to be decreased. This has to be done as the MultiNews dataset contains outlier news articles with a lot of sentences (see Figure 7). This can lead to an issue that paragraphs, which only contain a single sentence, are truncated and therefore possibly important information is lost. Due to the heterogeneity of the MultiNews dataset with regard to usage of paragraphs, truncating and padding with a fixed number of tokens will always lead to a trade-off between a lower batch size and an information loss. Both aspects can lead to a decreased model performance. Due to resource limitations, we adjusted the padding and truncation parameters in order to achieve a batch size of 3,000, for which the paragraph-level model achieved good results. Furthermore, sentence-level representation results in more vertices in the graph structure which leads to a higher memory consumption. As our results showed, an increased batchsize improves model performance on paragraph level. With limited resources, it is expected that a paragraph-level model outperforms a sentence-level model as it is possible to utilize larger batchsizes.

We were not able to train a sentence-level model on a batchsize of 4,000, which produced the best results for paragraph-level. It might be that a sentence-level model could outperform the paragraph-level model for a higher batchsize. However, preliminary results with different batchsizes showed that the paragraph-level model outperformed the sentence-level model for all configurations. As the memory consumption depends on the padding procedure as well as the batchsize, an increase of available resources also enables the possibility to increase the number of tokens per textual unit, which could counteract the issue of losing information by truncation.



**Figure 4: Visualization of correlation between ROUGE scores indicating source origin information via text similarity and the aggregated global attention weights for the decoding layer with the lowest correlation (left) and the decoding layer with the highest correlation (right) of the MultiNews dataset. A reduced test set is used for evaluation.**



**Figure 5: Visualization of correlation between ROUGE scores indicating source origin information via text similarity and the aggregated global attention weights for the decoding layer with the lowest correlation (left) and the decoding layer with the highest correlation (right) of the WikiSum dataset. A reduced test set is used for evaluation.**

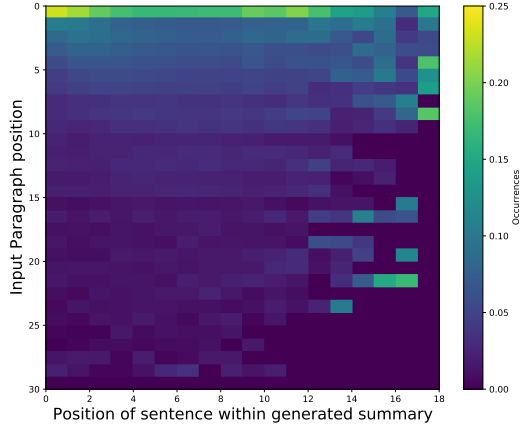
While for our comparisons we only constructed a similarity graph based on TF-IDF scores, graphs capturing lexical relations as well as discourse relations [2] are also shown to be beneficial for text summarization tasks [12]. However, the results of Li et al. were similar for all different graph representations. Thus, we omitted them here in our study. In addition, Christensen et al.’s discourse graph generally operates on token-level and captures the discourse relations between sentences and between not paragraphs [2]. Therefore, we did not use this graph type for our comparisons.

## 6.2 Source origin analysis

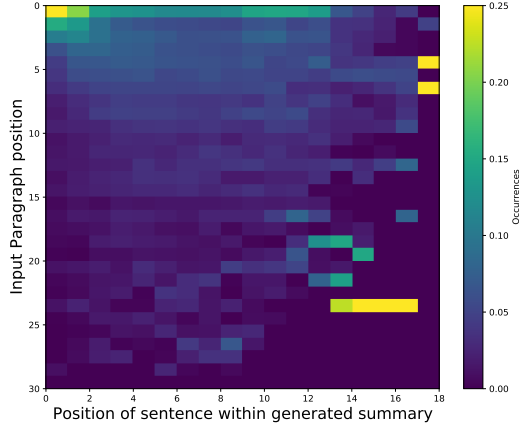
Our results on the source origin analysis using the attention weights show that the multi-heads in a single decoding layer attend mostly

the same paragraphs. This is indicated by a high correlation between the attention weights. We argue that the multi-heads are still able to extract different information while attending the same paragraphs, since the output of a multi-head attention mechanism depends on trainable parameters, which extract information based on the attended input paragraphs. Our experiments show that the AWD can be used to indicate which paragraphs provided relevant information for the generated summaries, especially the attention weights of deeper layers are suitable for this task. This was expected because the output of the transformer layer is a linear transformed weighted average of the input paragraphs as shown in Equation 1. Furthermore, in deep neural networks for computer vision, it is known that early layers learn to extract high level features while layers closer to the output extract more task specific features [22].





(a) Decoding Layer 3



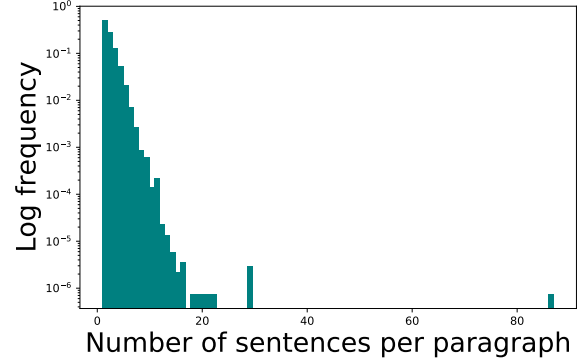
(b) Decoding Layer 6

**Figure 6: Heatmap to visualize the paragraph positions, which were attended the most for each generated sentence. The analysis is performed on the MutliNews dataset.**

Similar effects may apply to language models, as indicated by the success of transfer-learning and fine-tuning of BERT for language processing tasks [3].

Even though we expected a gradually change for the correlation of consecutive layers, it is particularly noticeable that the decoding layers can be clustered into two sets with regard to the similarity of their attention weights. There is a sudden jump of the correlations observed from layer 4 to 5 for the MultiNews dataset and 5 to 6 for the WikiSum dataset, see Table 2.

Based on the analysis of GraphSum’s attention mechanisms, our results show that the graph-based MDS model generates summaries mostly from the first paragraphs of each input document in the MultiNews dataset. Therefore, we assume that the GraphSum model



**Figure 7: Number of sentences per paragraph for the Multi-News dataset. Histogram is density normed and has a log-scaled y-axis to illustrate outliers.**

indeed learns the positional bias of the MultiNews dataset. However, it is not clear whether the model attends the first paragraphs based on their positional encoding or their actual content. Therefore it might be interesting to investigate how well a trained MDS model can perform on shuffled input documents, where the order of textual units (i. e., sentences, paragraphs or documents) is changed. This would allow to understand, if a trained model extracts information based on the position of a textual unit or its content. We leave this as part of future work.

We assumed that text similarity based on the ROUGE score can be used as a reference metric in our source origin analysis. One may argue that the ROUGE metric might be unsuitable as reference metric for abstractive summaries, which are heavily based on synonyms and/or paraphrasing, as the ROUGE metric indicates source origin information by calculating text overlaps between generated summaries and input paragraphs. However, abstractive summaries should still contain immutable key words (e. g., names, dates, locations) which can be captured by the ROUGE metric. Therefore, simple paraphrasing or the usage of synonyms for unimportant information should not distort the effectiveness of the reference metric.

An even better reference metric could be achieved if experts annotated the MultiNews dataset with source origin information manually. This labeled dataset could then be used to further evaluate the possibility of using attention weights to extract source origin information. Furthermore, to aggregate the token-level information of the attention weights to sentence-level, we calculated the mean value of the attention weights of the tokens. We assume that the mean is suitable for this task under the assumption that the model generates coherent summaries and therefore subsequent tokens should have a similar source origin. Other aggregation metrics like median could be evaluated in future work. Furthermore, the transformer architecture of the GraphSum model also calculates a local graph attention which attends input tokens. This local graph attention also directly influences the generated tokens. A combination of both attention weight distributions, i. e., global and local, could lead to even better result in terms of source origin information. In

general, we assume that the attention weights of any transformer based architecture in the MDS domain should be suitable for extracting source origin information. To verify this assumption, our approach could be applied to other transformer-based models.

## 7 CONCLUSION

Our experiments show that the summaries generated from a sentence-level representation of the MultiNews dataset using an extended GraphSum model are of similar quality than the paragraph-level representation. However, the sentence-level representation requires more computational resources as more vertices and edges are needed in the graph-based model. Thus, subsequently we investigated the source origin information by analyzing the attention weights on the paragraph-level representation. Here, we revealed that the attention weights of later decoding layers of the GraphSum’s decoder architecture contain valuable information about which input paragraphs provide the most information for the generated summaries. These weights in the MultiNews dataset correlate with the presence of a positional bias, i. e., that the most important information is located in the first sentences. Therefore, we conclude that the attention weights successfully learn a positional bias. This sheds some light on better explaining how abstract MDS using graph-based transformer models work. Note, the analysis of positional bias could not be done on the WikiSum dataset as it does not contain information about which paragraphs correspond to which input document. In future work, it will be interesting to investigate this property in the WikiSum dataset and other domains, as well as to apply our analyses to other types of transformer-based models.

The code of this work is available on GitHub: <https://github.com/arnelochner/GBTBMDS>.

## REFERENCES

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1409.0473>
- [2] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards Coherent Multi-Document Summarization. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, 1163–1173. <https://www.aclweb.org/anthology/N13-1136>
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [4] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22 (2004), 457–479.
- [5] 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. *arXiv:1906.01749* [cs.CL]
- [6] Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward. *arXiv:2005.01159* [cs.CL]
- [7] Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier Isn’t Always Better: Sub-aspect Analysis on Corpus and System Biases in Summarization. *arXiv:1908.11723* [cs.CL]
- [8] Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2019. Content Selection in Deep Learning Models of Summarization. *arXiv:1810.12343* [cs.CL]
- [9] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [10] Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226* (2018).
- [11] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Brussels, Belgium, 66–71. <https://doi.org/10.18653/v1/D18-2012>
- [12] Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging Graph to Improve Abstractive Multi-Document Summarization. *arXiv:2005.10043* [cs.CL]
- [13] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. <https://www.aclweb.org/anthology/W04-1013>
- [14] Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization. In *Proceedings of the 40th annual meeting of the association for computational linguistics*. 457–464.
- [15] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. *arXiv:1801.10198* [cs.CL]
- [16] Yang Liu and Mirella Lapata. 2019. Hierarchical Transformers for Multi-Document Summarization. *arXiv:1905.13164* [cs.CL]
- [17] Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* (2019).
- [18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762* [cs.CL]
- [19] Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous Graph Neural Networks for Extractive Document Summarization. *arXiv:2004.12393* [cs.CL]
- [20] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-Aware Neural Extractive Text Summarization. *arXiv:1910.14142* [cs.CL]
- [21] Michihiro Yasunaga, Rui Zhang, Kshitij Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based Neural Multi-Document Summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Association for Computational Linguistics, Vancouver, Canada, 452–462. <https://doi.org/10.18653/v1/K17-1045>
- [22] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How Transferable Are Features in Deep Neural Networks?. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (Montreal, Canada) (NIPS’14)*. MIT Press, Cambridge, MA, USA, 3320–3328.

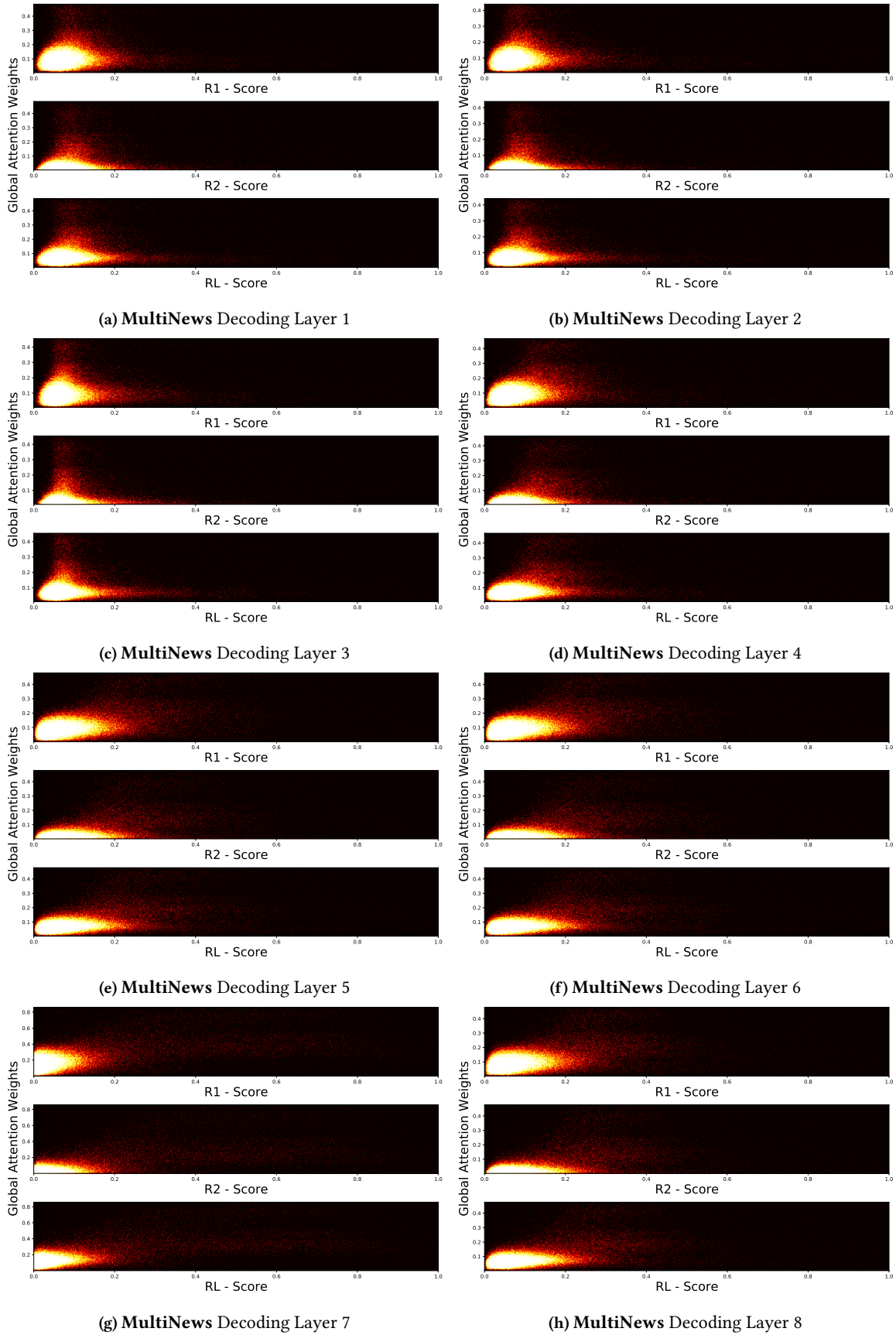
## **SUPPLEMENTARY MATERIALS**

### **A CORRELATION VISUALIZATION**

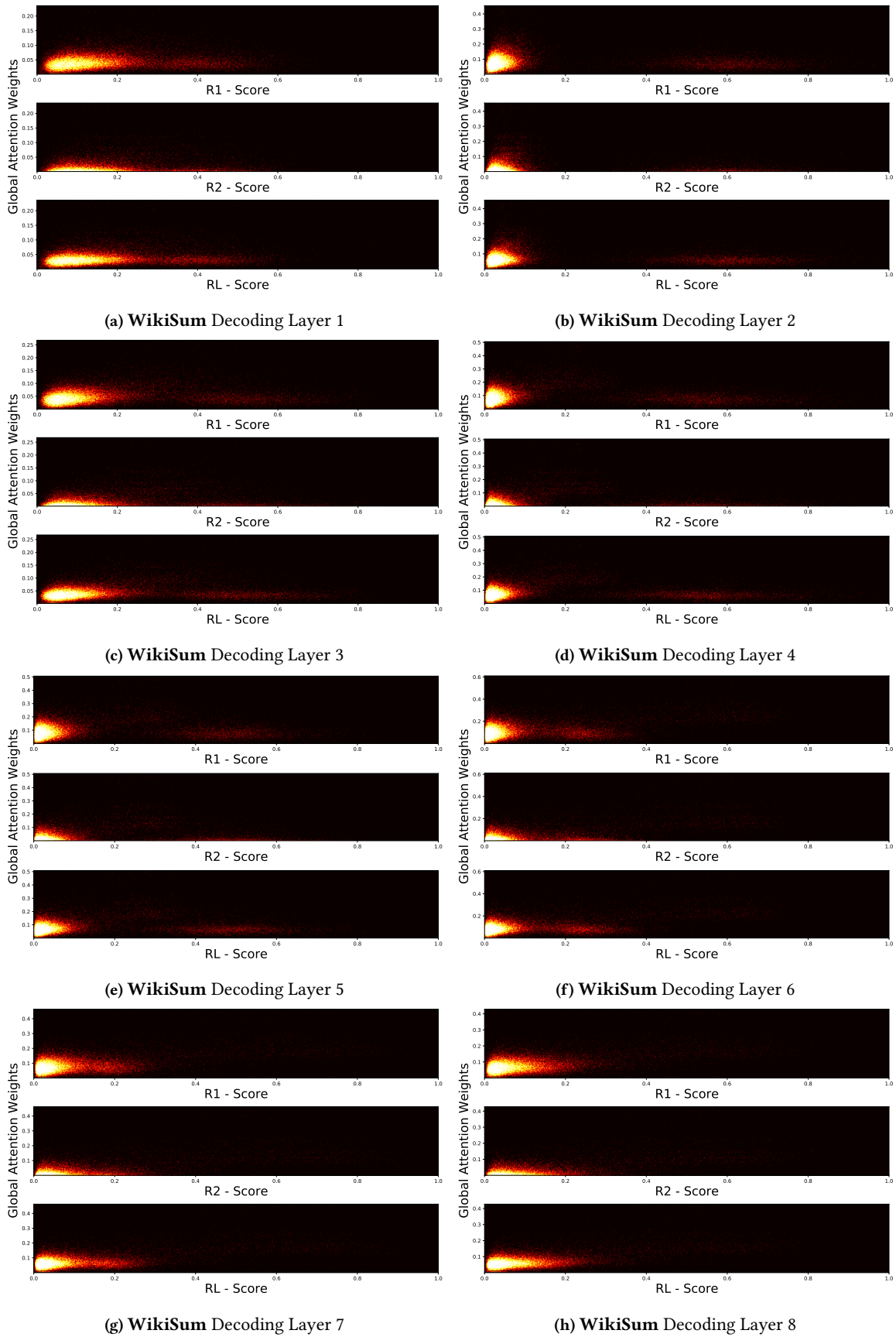
The correlation plots for the decoding layers, not visualized in Section 5.2 for MultiNews and WikiSum can be found in Figure 8 and Figure 9.

### **B POSITIONAL BIAS**

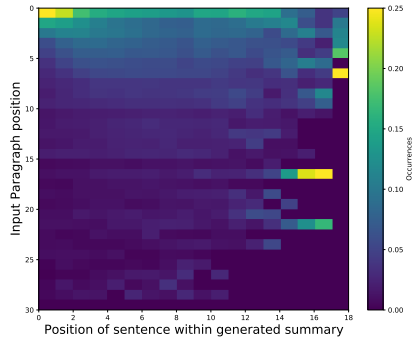
The presence of positional bias in the generated summaries indicated by attention weights is shown for all decoding layers in Figure 10.



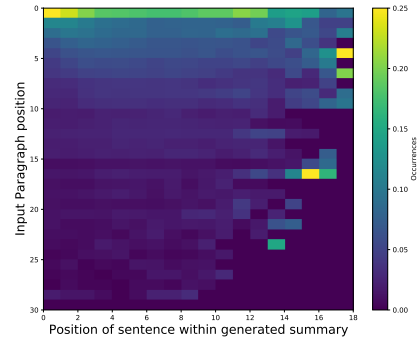
**Figure 8: Visualization of correlation between ROUGE scores indicating source origin information via text similarity and the aggregated global attention weights for the decoding layers not visualized in Figure 4. A reduced test set is used for evaluation.**



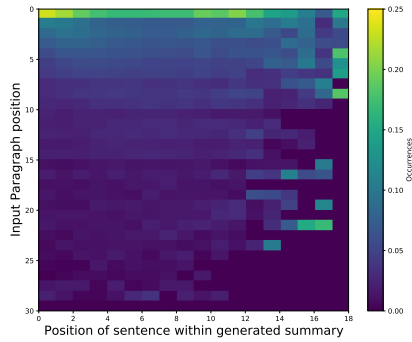
**Figure 9: Visualization of correlation between ROUGE scores indicating source origin information via text similarity and the aggregated global attention weights for the decoding layers not visualized in Figure 5. A reduced test set is used for evaluation.**



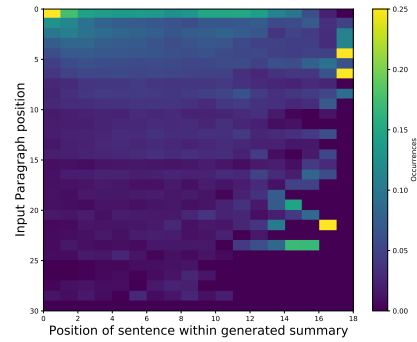
(a) Decoding Layer 1



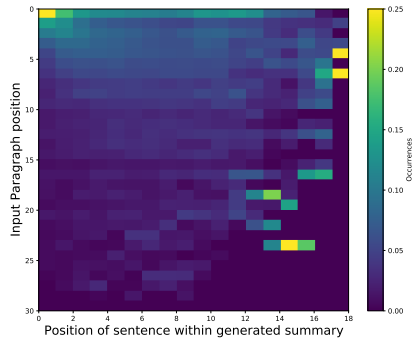
(b) Decoding Layer 2



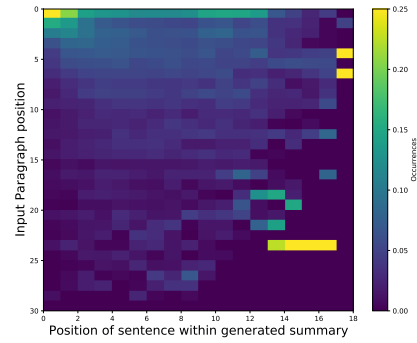
(c) Decoding Layer 3



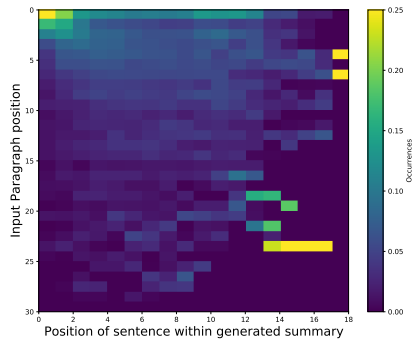
(d) Decoding Layer 4



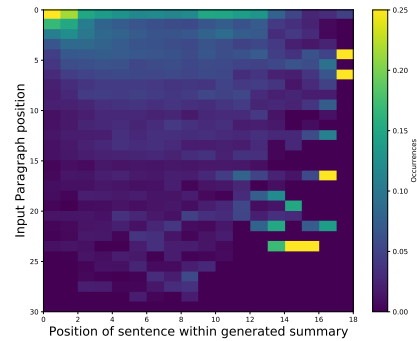
(e) Decoding Layer 5



(f) Decoding Layer 6



(g) Decoding Layer 7



(h) Decoding Layer 8

**Figure 10: Heatmap to visualize the paragraph positions, which were attended the most for each generated sentence. The analysis is performed on the MutliNews dataset.**