Modeling Endorsement for Multi-Document Abstractive Summarization

Anonymous EMNLP submission

Abstract

A crucial difference between single- and multi-document summarization is how salient content manifests itself in the document(s). While such content may appear at the beginning of a single document, essential information is frequently reiterated in a set of documents related to a particular topic, resulting in an endorsement effect that increases information salience. In this paper, we model the cross-document endorsement effect and its utilization in multiple document summarization. Our method generates a synopsis from each document, which serves as an endorser to identify salient content from other documents. Strongly endorsed text segments are used to enrich a neural encoder-decoder model to consolidate them into an abstractive summary. The method has a great potential to learn from fewer examples to identify salient content, which alleviates the need for costly retraining when the set of documents is dynamically adjusted. Through extensive experiments on benchmark multi-document summarization datasets, we demonstrate the effectiveness of our proposed method over strong published baselines. Finally, we shed light on future research directions and discuss broader challenges of this task using a case study.

1 Introduction

"Repeat a lie often enough and it becomes the truth." This proverb stresses the importance of *repetition* and *frequency* in human comprehension. It causes an endorsement effect that increases the salience of repeated information. In this paper, we leverage the endorsement effect to summarize multiple documents that discuss a particular event or topic (MDS). In the commercial arena, MDS could be used to aggregate search results (Miller, 2020) and distill insights from customer reviews (Bražinskas et al., 2020). Further, MDS is an integral part of the daily work of intelligence analysts who identify important information from raw documents and consolidate it into a summary report to be disseminated to the leadership (Hamilton, 2014).

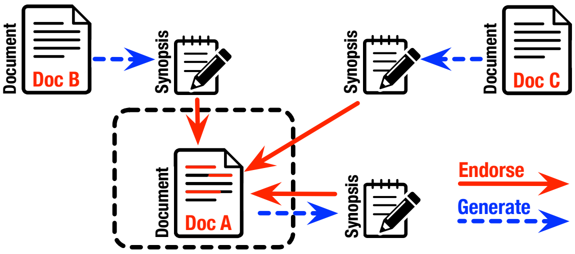

Figure 1: An example of synopsis-document relationships. Synopsis-document endorsements are leveraged to identify important text segments from a source document (e.g., Doc A). Strongly endorsed segments of all documents are consolidated into an abstractive summary.

Multi-document **A**bstractive **S**ummarization, i.e. **MuDAS**, remains a challenging problem compared to its single-document counterpart (See et al., 2017; Chen and Bansal, 2018; Narayan et al., 2018; Raffel et al., 2020; Lewis et al., 2020). The task poses a substantial challenge to modern neural models: when the set of source documents are concatenated into a flat sequence, it may exceed the maximum length allowed by the GPU memory. There is also fewer datasets available to train MuDAS models in an end-to-end fashion. Recent work tackles this problem by selecting representative sentences from the source documents to reduce the task to single-document summarization (Lebanoff et al., 2018; Coavoux et al., 2019; Fabbri et al., 2019).

Nevertheless, there could be substantial information loss if only representative sentences are used for MuDAS. It becomes unclear what information is reiterated and salient, resulting in unimportant sentence parts being included in the summary. E.g., when the sentence "*World leaders join to pledge $8 billion for vaccine, but the U.S. sits out*" is selected from the document set, it is unclear which of its segments, "*$8 billion*" or "*U.S. sits out*," is more salient given the topic of discussion. The neural representations also treat different quantities, e.g., "*$8 billion*" and "*$5 million*," indiscriminately (Rogers et al., 2020). Consequently, there is an urgent need for summarization systems to acquire fine-grained, segment-level textual salience. Without that, a neural abstractive system can miss out on salient details and favor fluency over information accuracy.

In this paper, we present a conceptual framework that leverages the endorsement effect to model fine-grained segment salience for multi-document summarization. When an analyst reads a document, he retains a synopsis of the key ideas of the document in his mind. The synopsis later serves as an endorser to identify segments in other documents that reiterate the same ideas (Hintzman, 1976). We call the synopsis an "**Endorser**" and the document a "**Candidate**." Segments of the candidate documents that are frequently endorsed by synopses suggest high salience and are to be consolidated into an abstractive summary. Our synopses are generated from a state-of-the-art summarizer (Lewis et al., 2020) and a variety of methods are investigated to quantify the level of endorsement from a text synopsis to a document. Figure 1 provides an overview of synopsis-document endorsement.

Our contributions in this paper include:

- presenting a new conceptual framework to model asynchronous endorsement from text synopses to documents for multi-document summarization;
- devising a novel method to enrich neural encoder-decoder models with fine-grained segment-level endorsement to consolidate strongly endorsed content into an abstractive summary; and
- through extensive experiments on multiple benchmark summarization datasets, we demonstrate the effectiveness of the endorsement method over state-of-the-art baselines.[1]

2 Related Work

Redundancy is essential in multi-document summarization. Without repetition and redundancy, even humans cannot agree on what information is salient and should be included in the summary (Daume III and Marcu, 2004). Optimizing summaries for frequency-based saliency has attained success prior to the era of deep learning (Berg-Kirkpatrick et al., 2011; Kulesza and Taskar, 2012; Boudin et al., 2015). These extractive systems strive to include the most frequently occurring concepts in the summary. However, when it comes to abstractive summarization systems, the frequency of concepts is not fully utilized by modern neural models.

Recent studies on MuDAS *implicitly* estimate frequency using hierarchical encoders / decoders. Liu and Lapata (2019) encode the documents using hierarchical Transformers where cross-document relationships are characterized by attention weights. Perez-Beltrachini et al. (2019) explore structured convolutional decoders. Li et al. (2020) leverage similarity and discourse graphs to alter the attention mechanism of encoder-decoder models. Despite promising progress, modeling frequency for multi-document summarization remains an open problem, in part because neural summarization models are often pretrained on single documents that contain little or no redundant content (Kryscinski et al., 2019; Zhang et al., 2019; Jin and Wan, 2020; Laban et al., 2020; Zhang et al., 2020a). Named entities and quantities that represent salient information details are not properly accounted for (Xu and Durrett, 2021). If we do not *explicitly* model frequency, abstractive summarizers may fail to adequately recognize such salient details.

We are particularly interested in reducing multiple input documents to a single document, then consolidate the content into a succinct abstract (Nayeem et al., 2018; Coavoux et al., 2019). Our method enhances the single document with fine-grained segment salience to offset the lead bias (Grenander et al., 2019; Xing et al., 2021), which hinders the development of multiple-document summarization. Our salience estimates are obtained from a frequency-driven endorsement model. Below we present details of the proposed method.

3 Summarization with Endorsement

We approach the MuDAS problem in two stages. First, we obtain fine-grained segment-level endorsement for any candidate document. By excluding unendorsed sentences from consideration, we reduce the set of documents to a single input document. We next present an enhanced abstractive summarization model to consolidate the document into a succinct abstract, analogously to how an editor would consolidate text with emphasis on endorsed segments. This process involves non-trivial design decisions. We start by presenting our summarization model with endorsement.

We choose the encoder-decoder architecture over decoder-only architectures (Radford et al., 2019; Dong et al., 2019; Brown et al., 2020). It allows us to balance the contribution from the source text and its endorsed segments in summary generation. The encoder and decoder each comprise of a stack of

[1] Our models and code will be made publicly available to enable future work to build on these results to develop robust multi-document abstractive summarizers.

L Transformer blocks (Vaswani et al., 2017). Let $\{x\}_{i=0}^{m}$ be the source sequence corresponding to the input document, and $\{y\}_{j=0}^{n}$ the summary sequence. x_0 and y_0 are beginning-of-sequence symbols. Let $\boldsymbol{E}$ be a matrix of token embeddings and $\boldsymbol{P}$ be position embeddings. An encoder produces a set of hidden vectors in its l-th layer (Eq. (1)), $\boldsymbol{H}^{(l)} = \langle \boldsymbol{h}_0^{(l)}, \ldots, \boldsymbol{h}_m^{(l)} \rangle$, where $\boldsymbol{h}_i^{(l)}$ is a hidden vector of the i-th source token. A decoder utilizes top-layer encoder hidden vectors $\boldsymbol{H}^{(L)}$ to decode the summary sequence, where $\boldsymbol{G}^{(l)}$ represents a sequence of hidden vectors of the l-th decoder layer (Eq. (2)). An upper triangular-shaped mask is used by the decoder, so that $\boldsymbol{g}_j^{(l)}$ only depends on summary tokens whose positions are less than j.

$$\boldsymbol{H}^{(l)} = \langle \boldsymbol{h}_0^{(l)}, \ldots, \boldsymbol{h}_m^{(l)} \rangle \tag{1}$$

$$= \begin{cases} \langle \boldsymbol{E}_{x_0} + \boldsymbol{P}_0, \ldots, \boldsymbol{E}_{x_m} + \boldsymbol{P}_m \rangle & l = 0 \\ \textsc{EncBlock}_l\big(\boldsymbol{H}^{(l-1)}\big) & l > 0 \end{cases}$$

$$\boldsymbol{G}^{(l)} = \langle \boldsymbol{g}_0^{(l)}, \ldots, \boldsymbol{g}_n^{(l)} \rangle \tag{2}$$

$$= \begin{cases} \langle \boldsymbol{E}_{y_0} + \boldsymbol{P}_0, \ldots, \boldsymbol{E}_{y_n} + \boldsymbol{P}_n \rangle & l = 0 \\ \textsc{DecBlock}_l\big(\boldsymbol{G}^{(l-1)}, \boldsymbol{H}^{(L)}\big) & l > 0 \end{cases}$$

With this architecture, we argue that it is preferable to modify the decoder and cross-attention to steer it towards endorsed content, rather than modifying the encoder representations $\boldsymbol{H}^{(L)}$, as they are often unsupervisedly pretrained. It would be best if such representations remain unaffected by whether a segment of the source text is endorsed or not to provide model flexibility. A decoder layer consists of three main blocks to transform from $\boldsymbol{G}^{(l-1)}$ to $\boldsymbol{G}^{(l)}$ (Eqs. (3-5)).[2] In particular, self-attention allows a summary token to attend to other summary tokens. Cross-attention allows a summary token to attend to all source tokens using $\boldsymbol{H}^{(L)}$. Finally, a feed-forward network with ReLU activation is applied to generate $\mathbf{G}^{(l)}$. Our focus of this work is to improve the cross-attention to emphasize on endorsed content during decoding.

$$\widetilde{\boldsymbol{G}}^{(l-1)} = \textsc{Self-Attn}\big(\boldsymbol{G}^{(l-1)}\big) \tag{3}$$

$$\widehat{\boldsymbol{G}}^{(l)} = \textsc{Cross-Attn}\big(\widetilde{\boldsymbol{G}}^{(l-1)}, \boldsymbol{H}^{(L)}\big) \tag{4}$$

$$\mathbf{G}^{(l)} = \textsc{FeedForward}\big(\widehat{\boldsymbol{G}}^{(l)}\big) \tag{5}$$

The *original* cross-attention head z transforms the j-th decoder state $\widetilde{\boldsymbol{g}}_j^{(l-1)}$ and i-th encoder state $\boldsymbol{h}_i^{(L)}$ into query, key and value vectors (Eqs. (6-8)). It computes attention weights as a normalized dot product between query and key vectors. The output of the head is a weighted sum of value vectors.

We introduce a set of *companion heads* for each original head. All companion heads of z share the parameters $\{\boldsymbol{W}_z^Q, \boldsymbol{W}_z^K, \boldsymbol{W}_z^V\}$, but a companion $\text{head}_j^{z,\tau}$ with an endorsement level of τ attends only to source tokens that are endorsed τ times or more. This is achieved with a special binary mask M_i^τ (Eqs. (9-10)). The original heads are believed to copy over source tokens that are deemed relevant to summary tokens according to the dependency syntax (Clark et al., 2019). The companion heads serve a similar purpose but have a narrower focus on endorsed source tokens—frequently endorsed tokens are more likely to be copied over by companion heads. The method thus improves head diversity similar to that of sparse Transformers (Correia et al., 2019; Huang et al., 2021). The hyperparameter τ controls the level of endorsement. Finally, all heads are pooled into a hidden vector $\widehat{\boldsymbol{g}}_j^{(l)}$ (Eq. (11)) to be passed to the feedforward layer.

$$\boldsymbol{q}_j^z = \boldsymbol{W}_z^Q \widetilde{\boldsymbol{g}}_j^{(l-1)} \qquad j \in [n] \tag{6}$$

$$\boldsymbol{k}_i^z = \boldsymbol{W}_z^K \boldsymbol{h}_i^{(L)} \qquad i \in [m] \tag{7}$$

$$\boldsymbol{v}_i^z = \boldsymbol{W}_z^V \boldsymbol{h}_i^{(L)} \qquad i \in [m] \tag{8}$$

$$\text{head}_j^{z,\tau} = \sum_{i=0}^{m} \frac{\exp(\boldsymbol{q}_j^{z\top} \boldsymbol{k}_i^z)}{\sum_{r=0}^{m} \exp(\boldsymbol{q}_j^{z\top} \boldsymbol{k}_r^z)} M_i^\tau \boldsymbol{v}_i^z \tag{9}$$

$$M_i^\tau = \begin{cases} 1 & \text{if Endorse}(x_i) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$\widehat{\boldsymbol{g}}_j^{(l)} = \sum_{z=1}^{n_{\text{head}}} \sum_{\tau=0}^{\tau_{\max}} \text{head}_j^{z,\tau} \boldsymbol{W}_z^\tau \tag{11}$$

When $\tau_{\max}$ is set to 0, the model reduces to its initial form using the original heads, i.e., $\text{head}_j^{z,0}$. Further, we initialize $\boldsymbol{W}_z^\tau = \lambda^\tau \boldsymbol{W}_z$, where $\boldsymbol{W}_z \in \mathbb{R}^{h_{\text{head}} \times h_{\text{model}}}$ are pretrained model parameters associated with the head z. $\lambda^\tau \in [0, 1]$ is a coefficient and $\boldsymbol{W}_z = \sum_{\tau=0}^{\tau_{\max}} \boldsymbol{W}_z^\tau$. It indicates that, head z and all of its companion heads are linearly interpolated to produce decoder hidden state $\widehat{\boldsymbol{g}}_j^{(l)}$. If a source token is not endorsed, it will have a reduced impact on the decoder hidden state when companion heads are used. The method has the advantage that, when new documents are dynamically added or removed from the set, it only changes the level of endorsement received by the tokens (τ), thus avoiding costly retraining of the neural encoder-decoder model. We proceed by describing how fine-grained segment-level endorsement is obtained from mod-

[2] We omit the residual connection and layer normalization associated with each block for brevity.

eling synopsis-document relationships.

4 Modelling Endorsement

Modelling endorsement serves two main purposes. It allows us to identify salient segments of text using a frequency-driven endorsement model, and the level of endorsement guides the summarizer to consolidate salient content. Further, it helps us reduce the source input from multiple documents to a single pseudo-document, whereby any unendorsed sentences are removed from consideration.

A fragment of text is considered to be endorsed if its information is observed in the endorser. We obtain a set of synopses from the source documents; they are used as *endorsers* to identify salient segments from a candidate source document. A segment that is endorsed only once indicates its information is considered important by only one source document. Frequent endorsement by multiple endorsers suggests the information is reiterated in multiple source documents, and reiteration implies increased salience. Any information that is present among multiple sources is likely to be important. Thus, our method identifies salient segments considering both within- and cross-document saliency. Our approach is in spirit similar to those of building semantic concept graphs for multi-document summarization (Bing et al., 2015; Handler and O'Connor, 2018; Falke and Gurevych, 2019) in that frequently reiterated concepts are likely to be captured. However, we do not explicitly construct semantic concept graphs, but focus on modeling synopsis-document endorsement and incorporating it into summary generation, which distinguishes our work from these studies. We investigate two variants to compute segment-level endorsement.

4.1 Synopsis-Document Alignment

Let S be a synopsis serving as the endorser and D a source document, our goal is to estimate whether a token x_i of the document is endorsed by the synopsis. A soft alignment between the synopsis and document is attainable by utilizing text evaluation metrics such as BERTScore (Zhang et al., 2020b), where we build contextualized embeddings for tokens of the document and synopsis, compute the cosine similarity of embeddings, and find a most similar synopsis token for each token of the document to obtain the endorsement score $\mathcal{S}(x_i)$ (Eq. (12)). Albeit a greedy alignment, the method can produce competitive results comparing to methods such as the earth mover's distance (Zhao et al., 2019).

$$\mathcal{S}(x_i) = \max_{y_j \in S} \text{Sim}(x_i, y_j) \tag{12}$$

Contiguous Segments It is important to endorse segments of text rather than isolated tokens, as segments such as "*$8 million*" is either included in the abstract in its entirety, or not at all. We transform token-level endorsement scores into binary decisions using the maximum sum subarray algorithm (Eq. (13)), which finds a contiguous subsequence that yields the highest sum of scores. The solution is trivial when all scores are positive. We thus offset the scores by δ before applying the algorithm. Let $\{0.2, 0.3, -0.1, 0.4, -0.5\}$ be an example of a set of adjusted endorsement scores, the algorithm endorses the first four tokens as the sum of their scores is the highest, yielding $\{1, 1, 1, 1, 0\}$, where 1 indicates the token is endorsed and 0 otherwise. We apply the algorithm to each sentence of the document and discard the segment if it has less than 5 tokens. The method endorses salient segments of text, yet is lenient to include gap tokens.

$$\{s, e\} = \arg\max_{\{i,j\} \in m} \sum_{k=i}^{j} (\mathcal{S}(x_k) - \delta) \tag{13}$$

Soft vs. Hard Alignment A hard alignment between the synopsis and document can be obtained from string matching. A document token receives a score of 1 if it finds a match in the synopsis. Similar to above, we offset the scores by δ to obtain segments of endorsed text. Hard alignment is sensitive to entities and quantities; yet it can miss out on paraphrases. We compare the effectiveness of these alignment methods in the results section.

4.2 Synopses as Endorsers

A synopsis contains the main points of the source document. We employ BART (Lewis et al., 2020) as a single-document abstractive summarizer to produce a synopsis from each document of the input cluster. Synopses as endorsers are superior to whole documents or sentence extracts. Not only are synopses more concise, but they can exclude superfluous information such as quoted material from consideration. We score all sentences of the source documents according to the sum of their token endorsement scores. Highest endorsed sentences are selected and arranged in chronological order to form a pseudo-document, with a limit of $|D|$ tokens, which serves as the input to our summarization module.

When a token is deemed salient by τ endorsers, we set Endorse(x_i)=τ, analogous to a majority vote by the pool of endorsers. We introduce *reciprocal endorsement*, where a synopsis can endorse every document of the cluster; and *sequential endorsement*, where source documents are arranged in chronological order and only synopses of the later documents can endorse the earlier documents. Sequential endorsement assumes the first few articles of an event or topic are more important than others. It avoids endorsing redundant content, which is particularly useful when the documents contain redundancy or noise that is typical in the output of clustering algorithms for content aggregation. Importantly, our endorsement framework offers a potential to customize endorsement patterns based on the trustworthiness of news sources, political leanings, content quality, and more.

5 Data

We experiment with a large-scale multi-document summarization dataset (Gholipour Ghalandari et al., 2020) whose data are gathered from the Wikipedia Current Events Portal (WCEP).[3] The dataset contains an archive of important news events happening around 2016–2019. Each event is associated with a succinct summary of 30-40 words written by the editor and an average of 1.2 source articles linked from the event page. Additional source articles are retrieved from the CommonCrawl-News dataset using an event classifier. These articles are published within a window of ±1 day of the event date. We sample from these additional articles to ensure each event has 10 source articles. All summaries and source articles are in English. The dataset contains 8,158, 1,020 and 1,022 clusters respectively in the train, validation and test splits.

Our method aims to produce an abstractive summary from a cluster of news articles discussing a given event or topic. To assess the generality of our method, we apply the model trained on WCEP to three different test sets, i.e., the test split of WCEP and two benchmark multi-document summarization datasets, DUC-04 and TAC-11. The DUC/TAC datasets contain 50 and 44 clusters, respectively. They each comprise a set of news events collected over a period of time, and thus are suitable for evaluation of the model's generality in out-of-domain scenarios. DUC and TAC datasets contain four reference summaries per cluster created by NIST evaluators. WCEP has a single reference summary per cluster written by editors. The target summary length is 100 words for DUC/TAC and 40 words for WCEP, following the convention of previously published results. Endorsement-related statistics for these datasets are presented in Table 1.

Dataset	Synop Len	Num Segs	Seg Len	% Endorse Scores ≥ τ		
				$\tau=0$	$\tau=1$	$\tau=2$
WCEP	61	4.9	14.2	100.0	12.6	5.6
DUC-04	58	6.1	11.7	100.0	9.7	2.3
TAC-11	60	6.7	11.8	100.0	14.5	4.1

Table 1: (LEFT) The average length of synopses (SynopLen), average number of segments in a source document endorsed by a synopsis and average length of endorsed segments (SegLen). (RIGHT) Percentage of tokens with endorsement scores above the threshold value used in each set of companion heads. All tokens with scores below the threshold are masked out.

6 Experiments

Baseline Systems. We compare our endorsement method to strong multi-document summarization baselines. The extractive summarization systems include (i) *TextRank* (Mihalcea and Tarau, 2004) and *LexRank* (Erkan and Radev, 2004), which are graph-based approaches that perform strongly on this task. (ii) *Centroid* (Hong et al., 2014) computes the importance of a source sentence based on its cosine similarity with the document centroid. (iii) *Submodular* (Lin and Bilmes, 2011) treats multi-document summarization as a submodular maximization problem. (iv) *KL-Sum* (Haghighi and Vanderwende, 2009) is a greedy approach that adds sentences to the summary to minimize KL divergence. (v) *TSR* and *BertReg* (Gholipour Ghalandari et al., 2020) are regression-based sentence ranking methods using averaged word embeddings (TSR) and BERT sentence embeddings (BertReg).

The abstractive summarization systems include: (vi) *PointerGen* (See et al., 2017), which generates a summary by copying source words and predicting new words. The set of documents are concatenated to form the input. (vii) *PG-MMR* (Lebanoff et al., 2018) exploits the maximal marginal relevance method to select sentences and an encoder-decoder model to fuse them into an abstract. (viii) *Hi-MAP* (Fabbri et al., 2019) introduces an end-to-end hierarchical attention model to generate abstracts from multi-document inputs. We compare our system to these baselines and report results on WCEP, DUC-04, and TAC-11 datasets[4].

[3] https://en.wikipedia.org/wiki/Portal:Current_events

[4] We were unable to compare our method with hierarchical Transformers (Liu and Lapata, 2019) because the authors did not make their ranker available for ranking paragraphs.

Sequential vs. Reciprocal Endorsement. We investigate two endorsement patterns: (a) *reciprocal endorsement* allows any two documents of the same cluster to endorse each other, and (b) *sequential endorsement* arranges source documents in chronological order and only later documents are allowed to endorse earlier ones. The endorsement mechanism provides the flexibility needed for many domains to exploit cross-document relationships to generate abstractive summaries. For our variants, the highest-scoring sentences are consolidated to form an input document which, along with the endorsement scores, are passed to our endorsement-aware abstractor to be condensed into a summary.

Endorsement-Aware Abstractor. We employ BART, a state-of-the-art encoder-decoder model as our base abstractor (Lewis et al., 2020). The model has 24 layers in the encoder and decoder, a hidden size of 1024, 16 heads, with a total of 406M parameters. It was fine-tuned on the train split of WCEP for an average of two epochs with a batch size of 4. We use the Adam optimizer (Kingma and Ba, 2015) and a learning rate of 3^{-5} with warm-up. At inference time, we use a beam size of K=4, with a minimum decoding length of 10 and a maximum of 50 tokens. Our implementation is based on fairseq[5] and it takes about two hours to train the model on a NVIDIA V100 32GB GPU card.

For the endorsement-aware abstractor, we add two sets of companion heads to the decoder, for a total of 48 attention heads. The τ values for each set of heads are 0/1/2. Table 1 shows the percentage of tokens that receive different levels of attention: 12% of the tokens receive level-1 attention ($\tau = 1$), 4% receive level-2 attention ($\tau = 2$). The λ^τ values are set to be 0.8, 0.1, and 0.1—this gives more influence to the original attention heads, so the model is not confused by the addition of the new heads that attend to endorsed segments. We use a maximum of 1024 tokens for the input document.

Synopsis-Document Endorsement. To enable soft alignment between a synopsis and a candidate document, we use BERTScore (Zhang et al., 2020b) with the following hash code: roberta-large_L17_no-idf_version=0.3.2(hug_trans=2.8.0)-rescaled. It suggests that the token representations are drawn from the 17th layer of RoBERTa-large. Our maximum sum subarray algorithm requires the scores to contain a mix of positive/negative values. Thus, we subtract all scores by δ. The δ values are 0.85 and 0.8 for the soft and hard alignment, respectively. These values are tuned on validation data, where a larger δ indicates fewer tokens will be endorsed.

System	R-1	R-2	R-SU4
Extractive			
Random Lead	27.6	9.1	–
Random	18.1	3.0	–
TextRank	34.1	13.1	–
Centroid	34.1	13.3	–
Submodular	34.4	13.1	–
TSR	35.3	13.7	–
BertReg	35.0	13.5	–
Our Method (In-Domain)			
Endorser-Reciprocal	43.3	21.9	22.1
Endorser-Sequential	**45.4**	**23.2**	**23.5**

Table 2: A comparison of multi-document summarizers on WCEP's test set.[6] *Endorser-** are our proposed methods.

We proceed by presenting summarization results on our datasets, including an ablation study to examine the contribution of each part of our method. We also present a case study showcasing the potential of our endorsement method.

6.1 Results

Our methods achieve state-of-the-art results when compared to previous work on WCEP's test set (Table 2). Sequential endorsement outperforms reciprocal endorsement due to the ability of sequential endorsement to remove redundancies introduced in later documents. In news domain, later articles generally review information from previous articles and introduce small developments in the story. By ordering the documents chronologically and having later articles give endorsement to earlier articles, it encourages the summarizer to pick content from earlier articles and reduce redundancy introduced in later articles. The largest performance increase can be seen in R-2, with *Endorser-Sequential* achieving a 9.7 increase over a BERT-based method. It demonstrates the effectiveness of endorsement for detecting salient segments and stitching them together to form a summary.

We report experimental results on DUC-04 and TAC-11 datasets in Tables 3 and 5. Here, our methods can outperform or perform comparably to previous summarization methods. On the WCEP test set, it corresponds to an ***in-domain*** scenario. On DUC-04 and TAC-11 test sets, it is an ***out-of-domain*** scenario. Due to data scarcity, the model can only be

[5] https://github.com/pytorch/fairseq

[6] We note that baseline summarizers use a maximum of 100 articles per cluster; these results are obtained from Gholipour Ghalandari et al. (2020). In contrast, our endorsement methods outperform the baselines with only 10 input articles per cluster.

System	R-1	R-2	R-SU4
Extractive			
TextRank	33.16	6.13	10.16
LexRank	34.44	7.11	11.19
Centroid	35.49	7.80	12.02
Neural Abstractive			
Pointer-Gen	31.43	6.03	10.01
PG-MMR	**36.88**	8.73	12.64
PG-BRNN	29.47	6.77	7.56
Hi-MAP	35.78	8.90	11.43
Our Method (Out-of-Domain)			
Endorser-Sequential	34.74	8.08	12.06
Endorser-Reciprocal	35.24	8.61	12.49
Endorser-Oracle	36.27	**8.93**	**13.04**

Table 3: A comparison of multi-document summarizers on the DUC-04 dataset. *Endorser-** are our methods.

System	R-1	R-2	R-SU4
Endorser-HardAlign	44.7	22.4	22.6
Endorser-SoftAlign	45.4	23.2	23.5
- companion heads	**45.8**	**23.5**	**23.8**
- endorse selection	43.6	23.0	22.9
- abstractive module	28.3	9.3	10.9

Table 4: Ablation study on WCEP dataset.

System	R-1	R-2	R-SU4
Extractive			
KLSumm	31.23	7.07	10.56
LexRank	33.10	7.50	11.13
Neural Abstractive			
Pointer-Gen	31.44	6.40	10.20
PG-MMR	37.17	10.92	14.04
Our Method (Out-of-Domain)			
Endorser-Sequential	36.11	9.52	13.07
Endorser-Reciprocal	37.43	10.71	13.94
Endorser-Oracle	**38.01**	**11.11**	**14.61**

Table 5: A comparison of multi-document summarizers on the TAC-11 test set. *Endorser-** are our methods.

trained on the train split of WCEP and then tested on DUC/TAC datasets. The fact that our system, when used out-of-the-box, can attain better or comparable results to the previous state-of-the-art has demonstrated its strong generalization capability. It suggests that obtaining segment-level endorsement on an outside domain then using it to inform summary generation is meaningful.

We observe that the reciprocal endorsement strategy outperforms sequential endorsement for DUC-04 and TAC-11 test sets. A closer look at the data suggests that this is due to the lower amount of redundancy present in DUC/TAC data. While WCEP documents are *automatically clustered* and contain much redundancy, source documents of DUC/TAC are *manually selected* by NIST assessors, each successive document in a topic cluster presents new developments about the topic. Thus, reciprocal endorsement may lead to better results for domains with less redundancy.

Intuitively, we want to steer the model attention towards endorsed segments if they are of high quality, and away from the segments otherwise. We conduct a set of oracle experiments that set λ^{τ} values to be proportional to the R-2 recall scores of endorsed segments. If the segments obtained for $\tau = 2$ yield a high R-2 recall score, they contain summary content and the model should attend to these endorsed segments by using a high λ^{τ} value. Results are reported in Tables 3 and 5 (*Endorser-Oracle*). We find that such a strategy is effective for making the most of companion heads. Future work may associate attention (λ^{τ} values) with the quality of segments obtained at different levels of endorsement ($\tau = \{0, 1, 2\}$).

6.2 Ablation

We perform an ablation study on WCEP to study the effects of each component in our model (Table 4). First, we compare the endorsement methods, denoted by *HardAlign* and *SoftAlign*. SoftAlign achieves consistently better results, showing that it is important to allow flexibility when aligning synopses to documents for endorsement. Next, we remove several components from the best-performing model (SoftAlign) to understand the effect of each. Removing "companion heads" from the abstractive model results in a very small boost in performance. Removing "endorsement selection"—meaning the model uses no information gained from performing endorsement, and is simply a BART model trained to summarize documents—leads to a significant performance drop, especially in R-1. It suggests that using endorsement to identify summary-worthy content from multiple documents is beneficial for an abstractive model.

Moreover, removing the "abstractive model"—meaning summaries are created extractively by selecting the highest-endorsed sentences—results in a large decrease in scores. It indicates that content-selection by endorsement cannot be done alone without an abstractor to create a more concise summary. This is especially the case for WCEP, where human reference summaries are relatively short.

We additionally report BERTScore (Zhang et al., 2020b) to evaluate summaries, in addition to the ROUGE metric (Lin, 2004). BERTScore uses cosine similarity between BERT contextual embeddings of words to detect word overlap between two texts, thus overcoming the problem of lexical variation in summarization. On DUC-04, the F_1 scores

(a) Single Synopsis Generated by BART

Opposition leader Sam Rainsy seeks clarification of security guarantees promised by Hun Sen. Hun Sen announced a government guarantee of all politicians' safety Wednesday. The opposition leader was forced to take refuge in a U.N. office in September to avoid arrest. The two parties have formed three working groups to hammer out details of the agreement.

(b) Endorsement from All Synopses

Sam Rainsy, who earlier called Hun Sen's statement "full of loopholes," asked Sihanouk for his help in obtaining a promise from Hun Sen that all members of the Sam Rainsy Party were free from prosecution for their political activities during and after last July's election. Sam Rainsy, a staunch critic of Hun Sen, was forced to take refuge in a U.N. office in September to avoid arrest after Hun Sen accused him of being behind a plot against his life. The alleged assassination attempt came during massive street demonstrations organized by the opposition after Hun Sen's Cambodian People's Party narrowly won the election. The opposition, alleging widespread fraud and intimidation, refused to accept the results of the polls. Fearing for their safety, Sam Rainsy and his then-ally Prince Norodom Ranariddh led an exodus of opposition lawmakers out of Cambodia after parliament was ceremonially opened in late September. Ranariddh, whose FUNCINPEC party finished a close second in the election, returned last week and struck a deal with Hun Sen to form a coalition government. The agreement will make Hun Sen prime minister and Ranariddh president of the National Assembly. The two parties have formed three working groups to hammer out details of the agreement, including the establishment of a Senate to be the upper house of parliament. Sok An, representing Hun Sen's party, said...

(c) Human-Chosen Segments

Sam Rainsy, who earlier called Hun Sen's statement "full of loopholes," asked Sihanouk for his help in obtaining a promise from Hun Sen that all members of the Sam Rainsy Party were free from prosecution for their political activities during and after last July's election. Sam Rainsy, a staunch critic of Hun Sen, was forced to take refuge in a U.N. office in September to avoid arrest after Hun Sen accused him of being behind a plot against his life. The alleged assassination attempt came during massive street demonstrations organized by the opposition after Hun Sen's Cambodian People's Party narrowly won the election. The opposition, alleging widespread fraud and intimidation, refused to accept the results of the polls. Fearing for their safety, Sam Rainsy and his then-ally Prince Norodom Ranariddh led an exodus of opposition lawmakers out of Cambodia after parliament was ceremonially opened in late September. Ranariddh, whose FUNCINPEC party finished a close second in the election, returned last week and struck a deal with Hun Sen to form a coalition government. The agreement will make Hun Sen prime minister and Ranariddh president of the National Assembly. The two parties have formed three working groups to hammer out details of the agreement, including the establishment of a Senate to be the upper house of parliament. Sok An, representing Hun Sen's party, said...

Table 6: An analysis of endorsed segments for a document. **(a)** A synopsis is generated from a candidate document. **(b)** The document also receives endorsement from the other 9 synopses in the cluster. **(c)** We compare to segments chosen by a human using the Pyramid method. Stronger highlighting indicates the segment received endorsement from many synopses.

are 29.89 and 30.14, respectively for our sequential and reciprocal model. The score for human reference summary is 35.08. They show very similar trends to those in Table 3, suggesting that our method when tested in out-of-domain scenarios can achieve competitive results.

6.3 A Case Study

We present an in-depth analysis of our fine-grained endorsement in Table 6. Soft alignment is used to endorse a candidate document from synopses of the cluster. We compare the resulting endorsements to the text segments chosen by a human using the Pyramid method (Nenkova and Passonneau, 2004), where semantic content units (SCUs) are identified from the reference summaries and are matched to phrases in the candidate document. The segments selected by our endorsement method and those chosen by manual annotation show a great amount of overlap, exemplifying the strength of our method in locating salient content from multi-document inputs. In fact, our endorsement method draws strong parallels with the Pyramid method—in our case, sentences from the automatically-generated synopses act as SCUs, which are then matched to phrases in the candidate document using a soft or hard alignment.

We observe that the endorsement given by a single synopsis is already quite similar to the human segments. However, taking the average endorsement from all ten synopses results in a higher quality set of segments. This shows the inherent value that exists from repetition in multi-document clusters, and it shows the importance of leveraging all of the documents rather than just a single one for salience estimation. Importantly, we observe that named entities, e.g., *"Sam Rainsy," "King Norodom Sihanouk,"* are more readily endorsed than other phrases. These entities are frequently repeated verbatim in all of the documents, thereby increasing their likelihood of being endorsed.

We envision future neural document summarization systems to produce better synopses than BART. They can lead to more accurate estimates for endorsed segments, hence improving the overall performance of our multi-document summarizer. The endorsement mechanism at its core is simple and robust—looking for shared content between a document and a synopsis. It provides great flexibility allowing the summarizer to potentially operate on document clusters containing a varying number of documents, which is a desirable characteristic.

7 Conclusion

We presented a novel framework to model asynchronous endorsement between synopses and documents for multi-document abstractive summarization. We introduced an endorsement method to enrich the encoder-decoder model with fine-grained endorsement. Our method was evaluated on benchmark multi-document datasets and we discussed challenges and shedded light on future research.

References

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 481–490, Portland, Oregon, USA. Association for Computational Linguistics.

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1587–1597, Beijing, China. Association for Computational Linguistics.

Florian Boudin, Hugo Mougard, and Benoit Favre. 2015. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1918, Lisbon, Portugal. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China. Association for Computational Linguistics.

Hal Daume III and Daniel Marcu. 2004. Generic sentence fusion is an ill-defined summarization task. In *Text Summarization Branches Out*, pages 96–103, Barcelona, Spain. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*.

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

Tobias Falke and Iryna Gurevych. 2019. Fast concept mention grouping for concept map-based multi-document summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 695–700, Minneapolis, Minnesota. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

Jillian Hamilton. A day in the life of an intelligence analyst [online]. 2014.

Abram Handler and Brendan O'Connor. 2018. Relational summarization for corpus analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1760–1769, New Orleans, Louisiana. Association for Computational Linguistics.

Douglas L. Hintzman. 1976. Repetition and memory. *Psychology of Learning and Motivation*, 10:47–91.

Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Hanqi Jin and Xiaojun Wan. 2020. Abstractive multi-document summarization via joint learning with single-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2545–2554, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Alex Kulesza and Ben Taskar. 2012. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xintong Li, Aleksandre Maskharashvili, Symon Jory Stevens-Guille, and Michael White. 2020. Leveraging large pretrained models for WebNLG 2020. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 117–124, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ron Miller. 2020. Former salesforce chief scientist announces new search engine to take on google. *https://techcrunch.com/2020/12/08/former-salesforce-chief-scientist-announces-new-search-engine-to-take-on-google/*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. 2019. Generating summaries with topic templates and structured convolutional decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5107–5116, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 948–954, Online. Association for Computational Linguistics.

Jiacheng Xu and Greg Durrett. 2021. Dissecting generation modes for abstractive summarization models via ablation and attribution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6925–6940, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.