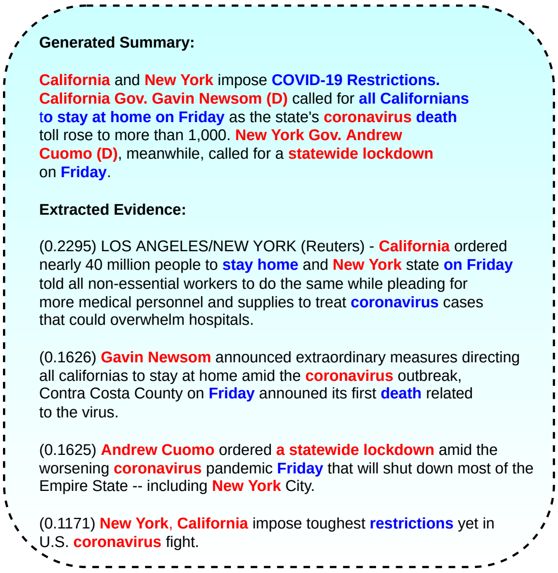
Multi-doc Hybrid Summarization via Salient Representation Learning

Min Xiao
Microsoft
mixia@microsoft.com

Abstract

Multi-document summarization is gaining more and more attention recently and serves as an invaluable tool to obtain key facts among a large information pool. In this paper, we proposed a *multi-doc hybrid summarization* approach, which simultaneously generates a human-readable summary and extracts corresponding key evidence giving a multi-doc input. To fulfill that purpose, we crafted a salient representation learning method to induce latent noteworthy features, which are effective for *joint* evidence extraction and summary generation. In order to train that model, we performed multi-task learning to optimize a composite loss, which is hierarchically constructed over the extractive and abstractive sub-components. We implemented such a fine system based on a ubiquitously-adopted transformer architecture and conducted experiments on a variety of datasets across two domains, achieving superior performance than the baselines.



Generated Summary:

California and **New York** impose **COVID-19 Restrictions.** **California Gov. Gavin Newsom (D)** called for **all Californians to stay at home on Friday** as the state's **coronavirus death** toll rose to more than 1,000. **New York Gov. Andrew Cuomo (D)**, meanwhile, called for a **statewide lockdown** on **Friday**.

Extracted Evidence:

(0.2295) LOS ANGELES/NEW YORK (Reuters) - **California** ordered nearly 40 million people to **stay home** and **New York** state **on Friday** told all non-essential workers to do the same while pleading for more medical personnel and supplies to treat **coronavirus** cases that could overwhelm hospitals.

(0.1626) **Gavin Newsom** announced extraordinary measures directing all californias to stay at home amid the **coronavirus** outbreak, Contra Costa County on **Friday** announed its first **death** related to the virus.

(0.1625) **Andrew Cuomo** ordered **a statewide lockdown** amid the worsening **coronavirus** pandemic **Friday** that will shut down most of the Empire State -- including **New York** City.

(0.1171) **New York**, **California** impose toughest **restrictions** yet in U.S. **coronavirus** fight.

Figure 1: The illustration of our hybrid system's outputs. The top rows present the generated summary and the bottom rows present the extracted key evidence (*aka.*, salient sentences). For each salient sentence, we also prepend the predicted salient score. This illustration demonstrates that the proposed novel system provides a *useful explainability mechanism* to the final summary.

1 Introduction

Multi-document summarization (MDS) aims to produce a concise summary of non-redundant salient facts based on multiple source documents under the same topic (Cao et al., 2017; Yasunaga et al., 2017), which is prevailingly useful in many application domains such as news-wire article summarization (Fabbri et al., 2019; Gu et al., 2020; Lee et al., 2022), scientific literature comparison (Lu et al., 2020; Shen et al., 2022a), civil rights lawsuits summarization (Shen et al., 2022b), and many others (Bražinskas et al., 2021; DeYoung et al., 2021).

Two main principled approaches are developed accordingly, multi-doc extractive methods (Cao et al., 2015a,b; Yin and Pei, 2015; Zhang et al., 2017) and multi-doc abstractive methods (Bražinskas et al., 2020; Amplayo and Lapata, 2021; Liu and Liu, 2021; Nan et al., 2021a; Pang et al., 2021). *Extractive multi-doc approaches* intend to directly extract non-redundant salient information from the original source (Mao et al., 2020; Wang et al., 2020; Parnell et al., 2022). It is usually carried with two stages, salience prediction and redundancy detection, where the former is trained with (pseudo-)salience labels (Cao et al., 2015b; Mao et al., 2020; Wang et al., 2020), and the latter is employed with ranking/selection tricks such as Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998; Zhang et al., 2017; Mao et al., 2020). Nonetheless, *abstractive multi-doc approaches* suggest paraphrasing the input to rewrite a smooth summary (Fabbri et al., 2019; Lewis et al., 2020; Ernst et al., 2022), where recent works are usually designed with dedicated components such as special attentions/encoders (Fabbri et al., 2019; Zhang et al.,

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 379–389
July 10-12, 2023 ©2023 Association for Computational Linguistics

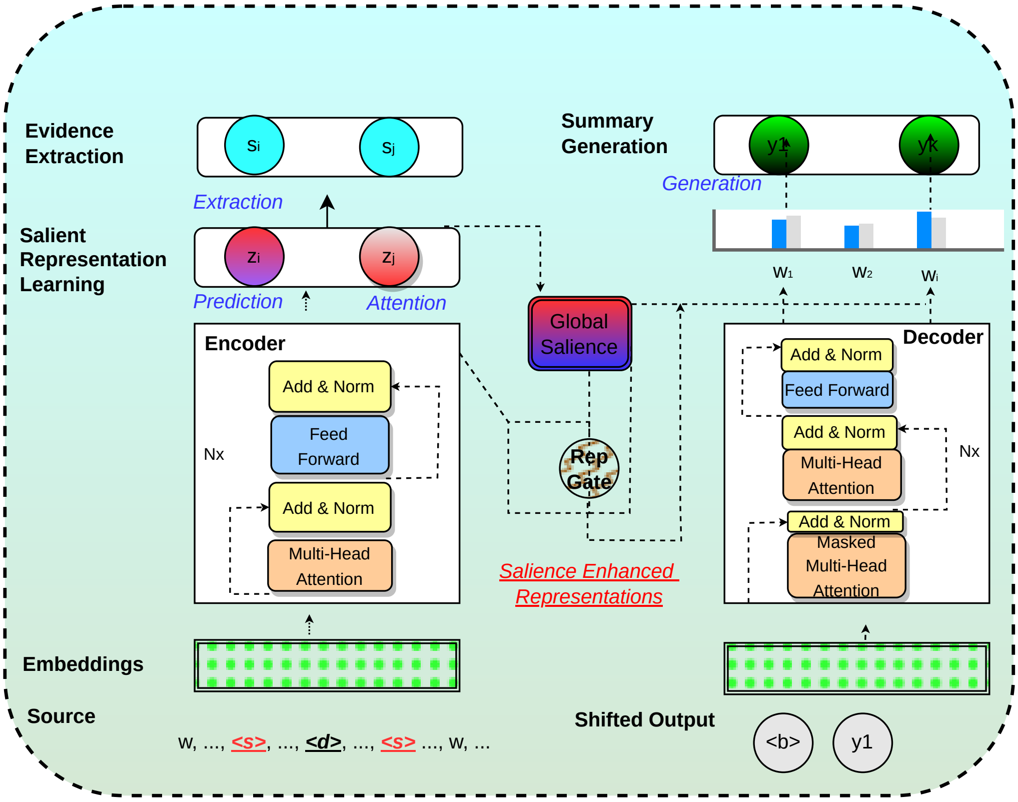


Figure 2: The architecture of the proposed multi-doc hybrid summarization system, powered by joint salient representation learning and multi-task training/prediction.

2020a; Pasunuru et al., 2021; Parnell et al., 2022; Song et al., 2022; Xiao et al., 2022).

In this paper, we advocate for a *hybrid* MDS system, which produces a paraphrased summary, achieved by the abstractive prediction module, attached with corresponding salient evidence, obtained from the *joint* extractive prediction module, where both modules are coupled and learned together, within an encoder-decoder-architectured transformer (Vaswani et al., 2017; Lewis et al., 2020; Xiao et al., 2022). *This is motivated from real-world business scenarios where enterprise customers are specifically requesting evidence extraction and alignment for abstractive summaries, which, as far as we know, was not fully supported in most commercialized AI platforms.* Please refer to Figure 1 for one such example to better understand it. In order to train such a hybrid system, we conducted multi-task learning to jointly optimize an extraction loss and generation loss. To align those two prediction tasks (*aka.*, summary generation and evidence extraction) appropriately, we exploited a salience attention with a gating mechanism, to effectively induce salient features. We present the overall architecture in Figure 2. In brief, given a multi-doc input, we feed it to our enhanced transformer for salient representation learning to jointly predict a final summary and extract notable evidence. Empirically, we carried extensive quantitative evaluations across multiple datasets, on various metrics such as ROUGE scores (Lin, 2004; Peng et al., 2021), perplexity (Jelinek et al., 1977), BERTScore (Zhang et al., 2020b), besides a manual qualitative evaluation with case studies (Novikova et al., 2017).

In summary, our contributions are twofold: 1) We proposed a novel multi-doc hybrid summarization system to generate linguistically-smooth summaries and to extract corresponding key evidence, based on the multi-doc inputs; 2) We conducted thorough empirical studies to quantitatively validate the effectiveness of the proposed approach and manually verify the quality of the extracted salient evidence.

2 Proposed Approach

Let $\mathbf{x}$ be a multi-doc input with n documents of the same topic, $\mathbf{x} = \{\mathbf{d_i}\}_{i=1}^{n}$, and each document be a *sequence* of sentences, *e.g.*, $\mathbf{d_i} = \left[\mathbf{s^i_1}, \ldots, \mathbf{s^i_j}, \ldots\right]$, where $\mathbf{s^i_j}$ is the j-th sentence of the i-th document $\mathbf{d_i}$. Similarly, each sentence is a *sequence* of words/tokens, *e.g.*, $\mathbf{s^i_j} = \left[\mathbf{w^{i,j}_1}, \ldots, \mathbf{w^{i,j}_k}, \ldots\right]$, where $\mathbf{w^{i,j}_k}$ indicates the k-th word/token of sentence $\mathbf{s^i_j}$. In this work, we are developing a prediction machine $\mathcal{M} : \mathcal{X} \to \mathcal{Y} \times \mathcal{S} \times \mathcal{Z}$ to map the input $\mathbf{x}$ to a triplet output such that $\mathcal{M}(\mathbf{x}) = \left(\mathbf{y}, \{\mathbf{s_{j'}}\}^q_{j'=1}, \{z_{j'}\}^q_{j'=1}\right)$, of which $\mathbf{y}$ is an abstractive summary, $\{\mathbf{s_{j'}}\}^q_{j'=1}$ is an evidence list of extracted q salient sentences, and $z_{j'}$ is the corresponding salient score for $\mathbf{s_{j'}}$.

Based on this hierarchical input, we first flatten the whole word sequences and insert a document-separator token, <d>, utilized to capture global interactions (Xiao et al., 2022), along with a sentence-separator token, <s>, used to assist sentence salience extraction. We then feed that flat sequence to our enhanced transformer (Figure 2), for joint salient representation learning and multi-task training/prediction.

2.1 Salient Representation Learning

Recently, encoder-decoder-based transformer has achieved great successes in the literature for a variety of generation tasks (Lewis et al., 2020), such as machine translation, summarization, and etc. Hence, our hybrid summarization system performed salient representation learning by enhancing such a transformer. Considering that we are tackling multi-doc inputs, where each instance (multiple documents) could be *much longer* than a single sentence or a short paragraph, we choose the longformer-encoder-decoder (LED) (Beltagy et al., 2020) as our backbone to implement such enhanced transformer, due to the fact that it has already demonstrated its capability to efficiently handle that longer sequences.

In general, our enhanced components consist of a *salient evidence extraction* module, which is constructed on top of the transformer encoder, and a *dynamic salience integration* module built across the encoder and decoder. Let $\mathbf{h_e} \in \mathbb{R}^{l \times m}$ be the hidden output of the transformer encoder, and $\mathcal{P} : \mathbb{R}^{l \times \cdot} \to \mathbb{R}^{l_s \times \cdot}$ be the sentence selection/filtering operator, where l denotes the whole sequence length and l_s denotes the number of sentence candidates. We first use a fully-connected feed forward (FF) layer with a dropout, followed with sentence candidate selection, to predict salience scores such that[1]

$$\hat{\mathbf{z}} = \mathcal{P}\left(\texttt{FF}\left(\texttt{Dropout}\left(\mathbf{h_e}\right)\right)\right) \in \mathbb{R}^{l_s}. \tag{1}$$

Then, we use the predicted salience scores $\hat{\mathbf{z}}$ to induce an attention weight for each sentence candidate and apply them to the sentences' hidden representations for self-aggregation. In particular, we adopt a salience attention to produce a *global salience* vector $\overline{\mathbf{h_s}}$ such that,

$$\overline{\mathbf{h_s}} = \sum_{j=1}^{l_s} a_j \mathcal{P}\left(\mathbf{h_e}\right)_j \in \mathbb{R}^m, \tag{2}$$

where $a_j = \frac{\exp(\hat{\mathbf{z}}_j)}{\sum_{j'} \exp(\hat{\mathbf{z}}_{j'})}$ is the attention weight for the j-th sentence.

Further, we connect this predicted global salience feature with decoder via dynamic integration during auto-regressive text generation. Specifically, we employ a gating mechanism (Zhou et al., 2016) to inject global salience features. Let $\mathbf{h_d} \in \mathbb{R}^m$ be the hidden output of the transformer decoder at one particular step, with this gating mechanism, the model is capable to dynamically select salience representations, such that,

$$\begin{aligned}
\mathbf{o_s} &= \overline{\mathbf{h_s}} \odot \mathbf{g} + \mathbf{h_d} \odot (1 - \mathbf{g}) \in \mathbb{R}^m, \\
\mathbf{g} &= \texttt{Sigmoid}\left(\texttt{FF}\left(\mathbf{h_g}\right)\right), \\
\mathbf{h_g} &= \texttt{GELU}\left(\texttt{Dropout}\left(\texttt{FF}\left(\overline{\mathbf{h_s}} \oplus \mathbf{h_d}\right)\right)\right),
\end{aligned} \tag{3}$$

where $\odot$ is element-wise product and $\oplus$ denotes concatenation, $\texttt{Sigmoid}(\cdot)$ and $\texttt{GELU}(\cdot)$ are the nonlinear activation functions (Hendrycks and Gimpel, 2016; Mishkin and Matas, 2016). We then use this dynamically gated salience representation $\mathbf{o_s}$ to generate a summary $\hat{\mathbf{y}}$. Note that the predicted salient scores $\hat{\mathbf{z}}$ will be sorted to retrieve salient sentences/evidence ($\hat{\mathbf{s}}$) during the inference stage.

2.2 Multi-Task Learning and Prediction

Due to the dual functionalities, we conduct a joint optimization wrt. the salient evidence extraction loss and the abstractive summary generation loss, such that

$$\mathcal{J} = \ell_1\left(\mathbf{y}, \hat{\mathbf{y}}\right) + \alpha \ell_2\left(\mathbf{z}, \hat{\mathbf{z}}\right), \tag{4}$$

where α is a trade-off parameter, $\ell_1(\cdot)$ denotes the abstractive loss, *e.g.*, token-level cross-entropy loss,

[1] For notation simplicity, we do not differentiate row and column vectors, assuming that the dimensionalities of them can be inferred from the context.

			Source			Target		Source → Target			
Domain	**Dataset**	**Samples**	Docs	Words	Sents	Words	Sents	Coverage	Density	Compression	**Ref**
News	Multi-News	56,216	2.7	2164.5	84.3	264.0	10.0	0.83	5.01	8.18	(Fabbri et al., 2019)

Table 1: The dataset statistics on the news domain, including document/sentence/word counts and source-to-target coverage/density/compression ratios.

		Test		**Lexical**			**Semantic**	**Fluency**	**Pred Summary**	
Domain	**Dataset**	**Samples**	**Method**	R-1_{f1} ↑	R-2_{f1} ↑	R-L_{f1} ↑	BS_{f1} ↑	PP↓	Words	Sents
News	Multi-News	5,622	Bart	49.75	20.00	24.76	35.61	13.22	263.4	9.6
			Primera	47.53	19.58	24.95	35.10	13.89	207.0	8.1
			MHS-SRL	**49.81**	**20.68**	**25.67**	**36.24**	**12.42**	252.4	9.8

Table 2: Empirical results on the news domain, where we underlined the numbers if the proposed system outputs the baselines, and used the **bold-font** to highlight the best number for each metric.

			Split		
Domain	**Dataset**	**Samples**	Train	Valid	Test
News	Multi-News	56,216	44,972	5,622	5,622

Table 3: The dataset-split stats on the news domain.

and $\ell_2(\cdot)$ denotes the extraction loss, *e.g.*, salience score error such as MSE.

As we can see, this paradigm of training requires labeled information for all candidate sentences. Hence, we employed a weakly supervised method to construct salience scores. Specially, we designed a hierarchical scoring method by first calculating the lexical coverages, *aka.*, ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), against the same gold summary and averaging them based on their F1 scores. Next, we computed a factuality-style score, NER-precision (Nan et al., 2021b) and averaged them to produce the pseudo salience labels for supervised training.

In summary, our systems share the similar functionality, respectively, with the extractive and abstractive method. Nevertheless, we performed them in a *joint hierarchical* way. Besides, our salience labels are more robust as we explored a composite method to combine lexical coverage measures (Lin, 2004) and a factuality-style measure (Nan et al., 2021b). Moreover, we used a dynamic gating mechanism to effectively integrate global salience during summary generation, enabling better alignment between two prediction tasks (*aka.*, summary generation vs. evidence extraction).

3 Experiments on the News Domain

In this section, we present empirical investigations for the proposed summarization system on news domain to justify its effectiveness.

3.1 Experimental Setup

The Multi-News dataset[2] (Fabbri et al., 2019) contains human-written summaries from the newser.com[3] site by professional editors, which are based on/linked with multiple source articles.We calculated important stats for this dataset in Table 1 and Table 3. In particular, we used the spaCy library[4] with the en_core_web_sm model to calculate word and sentence counts and used the Grusky et al. (2018)'s original tool to compute the coverage/density/compression ratios.

We compare our proposed system, **MHS-SRL** with two baselines: **Bart** (Lewis et al., 2020), which is widely used for summarization tasks, and **Primera** (Xiao et al., 2022), which is tailored to MDS systems. We set the batch size to be 8 and learning rate to be $3e^{-5}$, choose α from $\{0.1, 1, 10, 10^2\}$ based on the validation performance (dev/validation set), and used the Huggingface tool[5] for implementations.

3.2 Main Results

For each method, we finetuned it on the target dataset. We compare the proposed summarization system to the corresponding baselines with various metrics in Table 2, including the ROUGE scores (Lin, 2004) to estimate the lexical coverage, the

[2] https://github.com/Alex-Fabbri/Multi-News

[3] https://www.newser.com/

[4] https://spacy.io/

[5] https://huggingface.co

Gold Summary	– Kathryn Schurtz and Joseph Kearney were on their way to their wedding when the unthinkable happened Wednesday. The New Jersey couple was driving to Pittsburgh for the nuptials when a crash occurred ahead of them on I-78 in Windsor Township, Pa. Traffic slowed, but a tractor-trailer behind the couple wasn't able to stop in time and hit their car, pushing it into the back of another tractor-trailer. That crash caused a fiery chain reaction ultimately involving the couplesś car, which caught on fire, and five tractor-trailers, the Reading Eagle reports. Three other people were injured, NBC Philadelphia reports. Now, instead of a wedding, family will be attending funeral services. Schurtz, 35, "will be remembered for her voracious appetite for reading, love of cooking, and trailblazing new adventures with Joseph," says her obituary. Born and raised in New Jersey, she graduated from George Washington University before getting her MBA at Notre Dame. She was working for Oracle Data Cloud in New York City, and living back in New Jersey, when she was killed. Little information was available on her fiance, but NJ.com reports that per his Facebook profile, Kearney was originally from Pittsburgh. (This couple was married just one and a half hours before tragedy struck.)
Generated Summary	– A New Jersey couple was on their way to their dream wedding when they got caught up in a deadly chain-reaction crash on a busy Pennsylvania highway on the day before. Kathryn Schurtz, 35, and Joseph Kearney, 34, of Jersey City were driving to their wedding in Pittsburgh on Wednesday when they were hit from behind by a tractor-trailer and pushed into the back of another truck, which then pushed them into another truck and set them on fire, reports the Star-Ledger. The two were killed and three others were injured in the second crash on I-78 in Berks County, Pa., around 2:30pm, reports WNEP. Police say traffic was heavy in both westbound and eastbound lanes of the highway at the time of the crash, which occurred after an earlier crash. The crash slowed and stopped traffic for hours, and the highway was shut down for hours after the first crash. One lane of westbound traffic opened up around 7pm, but the two westbound lanes were still shut down as of Thursday morning. A family friend tells WFMZ that the couple, who had planned to marry in Pittsburgh, had been planning a wedding for a year and a half. "They were just so excited to get married," the friend says. "It was a dream come true." A GoFundMe page has been set up for the family, which has raised more than $10,000 so far.
Extracted Evidence	A New Jersey couple on their way to get married in Pittsburgh were killed in a fiery chain-reaction crash last week on Interstate 78 in Berks County, Pennsylvania. (0.1461)
	Kathryn Schurtz, 35, and her fiancé, Joseph Kearney, were driving on I-78 westbound in Windsor Township, Berks County, on Wednesday around 2:30 p.m. when the accident occurred. (0.1324)
	What to Know Loved ones are mourning a New Jersey couple who died in a chain reaction crash on I-78 in Berks County while driving to their wedding. (0.1251)
	Last week the couple was on their way to Pittsburgh the day before their dream wedding, when they got caught up in a tractor trailer crash in Berks County. (0.1205)
	Loved ones are mourning a New Jersey couple who died in a chain reaction crash on I-78 while on the way to their wedding. (0.1172)
	Kathryn M. Schurtz, 35, and Joseph D. Kearney, both of Jersey City, died in the collision Wednesday afternoon in Windsor Township. (0.1105)
	Kathryn Schurtz, 35, and her fiancé, Joseph Kearney, both died in the accident which involved their vehicle and five tractor trailers. (0.1099)
	Two people were killed and three were injured in a fiery chain-reaction crash on Interstate 78 in Windsor Township on Wednesday. (0.1080)
	It struck the couple's vehicle, which was then pushed into the back of another tractor-trailer and set off a chain reaction crash that included three more tractor-trailers. (0.1079)
	The couple was on the way to their wedding in Pittsburgh, Pennsylvania, at the time of the crash. (0.1035)

Table 4: Examples on the MULTI-NEWS dataset (1).

BERTScore (Zhang et al., 2020b) for semantic similarity[6], and the perplexity (Jelinek et al., 1977) for language smoothness. In addition, we present the word/sentence stats for the generated summaries for reference. As we can see, the proposed system performs much better than the corresponding baselines on all metrics by exploiting those effective salient features. Based on the *ablation study*, without the salient representation learning with the

[6] We used the debert model with baseline rescaling, which correlates better with human judgment, based on the authors' suggestions.

			Source			Target		Source → Target			
Domain	**Dataset**	**Samples**	Docs	Words	Sents	Words	Sents	Coverage	Density	Compression	**Ref**
Legal	Multi-LexSum$_{\text{D}\rightarrow\text{T}}$	1,603	10.7	125437.7	4591.0	24.7	1.4	0.85	1.88	5182.77	(Shen et al., 2022b)
	$_{\text{D}\rightarrow\text{S}}$	3,138	10.3	105255.2	3889.4	130.2	4.6	0.92	2.89	775.26	
	$_{\text{D}\rightarrow\text{L}}$	4,539	8.8	79632.5	2930.1	649.6	23.2	0.89	3.60	91.05	

Table 5: The dataset statistics on the legal domain, including document/sentence/word counts and the source-to-target coverage/density/compression ratios.

		Test		**Lexical**			**Semantic**	**Fluency**	**Pred Summary**	
Domain	**Dataset**	**Samples**	**Method**	R-1_{f1} ↑	R-2_{f1} ↑	R-L_{f1} ↑	BS$_{f1}$ ↑	PP↓	Words	Sents
Legal	Multi-LexSum$_{\text{D}\rightarrow\text{T}}$	312	Bart	25.39	8.18	20.79	29.50	20.36	27.6	1.4
			Primera	30.83	12.13	25.03	34.74	**15.45**	30.8	1.7
			MHS-SRL	**31.13**	**12.52**	**25.64**	**36.15**	20.33	25.2	1.6
	$_{\text{D}\rightarrow\text{S}}$	616	Bart	43.60	20.96	30.32	37.20	**7.80**	100.7	3.8
			Primera	46.88	23.06	**32.50**	**41.17**	8.77	102.5	4.0
			MHS-SRL	**47.72**	**23.56**	32.41	40.76	8.56	128.0	4.1
	$_{\text{D}\rightarrow\text{L}}$	908	Bart	48.13	22.84	28.21	37.14	12.17	366.2	12.2
			Primera	**53.52**	26.57	31.21	41.70	12.89	479.4	17.4
			MHS-SRL	53.25	**27.64**	**31.26**	**42.73**	**12.15**	479.4	17.4

Table 6: Empirical results on the legal domain. For each task, we underlined the numbers if the proposed system outperforms the baselines, and used the **bold-font** to highlight the best number for each metric.

			Split		
Domain	**Dataset**	**Samples**	Train	Valid	Test
Legal	Multi-LexSum$_{\text{D}\rightarrow\text{T}}$	1,603	1,130	161	312
	$_{\text{D}\rightarrow\text{S}}$	3,138	2,210	312	616
	$_{\text{D}\rightarrow\text{L}}$	4,539	3,177	454	908

Table 7: The dataset-split stats on the legal domain.

extraction loss, our system could drop the average ROUGE score by 4.26% and the BERTScore by 3.15%. To further justify that, we conduct a follow-on experiment by adding such salient representation learning with the extractive loss to the BART backbone, we observed a similar performance gain, *aka.*, uplifting the average ROUGE score by 2.49%, and the BERTScore by 1.57%.

3.3 Evidence Extraction vs. Summary Generation

Next, we focus on validating the extraction quality. Note that our system is able to produce a salience score for each sentence besides an abstractive summary. Hence, we conduct a qualitative evaluation by comparing the extracted evidence with the generated summary. From the test set, we randomly pick some examples for manual verification and present the results in Table 4. Note that the gold summary contains about 10 sentences on average (refer to Table 1). We then rank all the sentences based on the predicted salience scores and surface the top-10 as the evidence. To help the comparison, we also include the gold summary in Table 4. As we can see, the extracted evidence is indeed well aligned with the generated summary.

4 Experimental Results on the Legal Domain

In this section, we conduct experimental studies on the legal domain. The Multi-LexSum dataset[7] (Shen et al., 2022b) contains about 9K expert-authored summaries, distributed into three subsets (tiny, short, long), based on source documents of legal cases. Similarly, we present certain important stats in Table 5 and dataset-split stats in Table 7, and report the experimental results in Table 6. From Table 6, we can see that Primera method works much better than the Bart method due to the fact that the transformer architecture of Primera is more effective to handle much longer sequences (4096 vs 1024). Besides, another benefit comes from the task-specific pretraining vs. the task-agnostic pretraining, which is employed by the Bart method. For the MHS-SRL system, it could leverage the advantages of the Primera method in that both methods share a similar architecture. In addition, the dedicated salient representation learning, jointly trained with multi-task losses, brings in additional benefits. Given that our system is trained

[7] https://github.com/multilexsum/dataset

to score salience, we also quantitatively evaluated the extraction quality, which obtains the MSE score of $1.4e^{-4}$ on the Multi-LexSum$_{\mathrm{D}\rightarrow\mathrm{L}}$ subset, and $3.0e^{-4}$ for the Multi-LexSum$_{\mathrm{D}\rightarrow\mathrm{S}}$. All those empirical results validated the efficacy of joint salience extraction and summary generation.

Finally, we present the stats about the model parameter size and memory consumption in Table 8, which shows our hybrid approach is very efficient, which doesn't incur too much additional computing and memory resources.

Model	#Params	Est Mem Size
Bart	406 M	813 MB
Primera	447 M	894 MB
MHS-SRL	450 M	901 MB

Table 8: The stats of model parameters.

5 Conclusion

In this paper, we proposed to simultaneously perform summary generation and salient evidence extraction in a consistent way, where extracted salient representations are directly employed to enhance abstractive summary paraphrasing. To validate the efficacy, we conducted thorough experimental studies over multiple datasets across different domains. Both quantitative and qualitative results reveal the effectiveness of this novel summarization system.

Acknowledgements

We would like to thank all reviewers for their constructive comments and suggestions, which helped us to improve the quality of this manuscript. We sincerely appreciate their time and efforts.

References

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and controllable opinion summarization. In *EACL*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: the long-document transformer. *arXiv*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *EMNLP*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *EMNLP*.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2017. Improving multi-document summarization via text classification. In *AAAI*.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015a. Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*.

Ziqiang Cao, Furu Wei, Sujian Li, Wenjie Li, Ming Zhou, and Houfeng Wang. 2015b. Learning Summary Prior Representation for Extractive Summarization. In *ACL-IJCNLP*.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: A dataset for multi-document summarization of medical studies. In *EMNLP*.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2022. Proposition-Level Clustering for Multi-Document Summarization. In *NAACL*.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *ACL*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *NAACL*.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, You Wu, Cong Yu, Daniel Finnie, Hongkun Yu, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating Representative Headlines for News Stories. In *WWW*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv*.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *NAACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. In *ACL-IJCNLP*.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A Large-scale Dataset for Extreme Multi-document Summarization of Scientific Articles. In *EMNLP*.

Yuning Mao, Yanru Qu, Yiqing Xie, Xiang Ren, and Jiawei Han. 2020. Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In *EMNLP*.

Dmytro Mishkin and Jiri Matas. 2016. All you need is a good init. In *ICLR*.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021a. Improving factual consistency of abstractive summarization via question answering. In *ACL-IJCNLP*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021b. Entity-level factual consistency of abstractive text summarization. In *EACL*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *EMNLP*.

Richard Yuanzhe Pang, Adam D. Lelkes, Vinh Q. Tran, and Cong Yu. 2021. AgreeSum: Agreement-oriented multi-document summarization. In *Findings of the ACL-IJCNLP*.

Jacob Parnell, Inigo Jauregi Unanue, and Massimo Piccardi. 2022. A multi-document coverage reward for relaxed multi-document summarization. In *ACL*.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently Summarizing Text and Graph Encodings of Multi-Document Clusters. In *NAACL*.

Xutan Peng, Yi Zheng, Chenghua Lin, and Advaith Siddharthan. 2021. Summarising Historical Text in Modern Languages. In *EACL*.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022a. MReD: a meta-review dataset for structure-controllable text generation. In *Findings of the ACL*.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022b. Multi-LexSum: real-world summaries of civil rights lawsuits at multiple granularities. In *NeurIPS*.

Yun-Zhu Song, Yi-Syuan Chen, and Hong-Han Shuai. 2022. Improving Multi-Document Summarization through Referenced Flexible Extraction with Credit-Awareness. In *NAACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NIPS*.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous Graph Neural Networks for Extractive Document Summarization. In *ACL*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization. In *ACL*.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based Neural Multi-Document Summarization. In *CoNLL*.

Wenpeng Yin and Yulong Pei. 2015. Optimizing sentence modeling and selection for document summarization. In *IJCAI*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Yong Zhang, Meng Joo Er, Rui Zhao, and Mahardhika Pratama. 2017. Multiview Convolutional Neural Networks for Multidocument Extractive Summarization. *IEEE Transactions on Cybernetics*, 47(10):3230–3242.

Guo-Bing Zhou, Jianxin Wu, Chen-Lin Zhang, and Zhi-Hua Zhou. 2016. Minimal gated unit for recurrent neural networks. *International Journal of Automation and Computing*, 13:226–234.

A More examples on the MULTI-NEWS dataset

In Table 9 and Table 10, we provide additional examples.

Gold Summary	– Madonna doesn’t mind criticism of her new movie, WE—a good thing, since it bombed at the Venice Film Festival and, Reuters notes, subsequently received a one-star review in the Guardian—as long as that criticism is directed at the movie, and not at the Material Girl. "I can tell when people are reviewing my film and when they’re reviewing me personally," Madonna said yesterday at the Toronto International Film Festival, where WE is screening. "So when they stick to the film, then I do care" what critics think, she explained. She also described King Edward VIII, whose romance with Wallis Simpson is depicted in the film, as “very punk rock,” the Telegraph reports. That’s why she put the Sex Pistols’ “God Save the Queen” on the soundtrack, she explained: “I thought [the king] was quite rebellious and cutting edge in his point of view about life and about how to run the empire and using the Sex Pistols was a perfect marriage.” (She also, the Telegraph adds, broke into song at one point during the Q&A session—watch at left.) Click for another story coming out of the film festival—apparently, Madge doesnt́ like to be looked at by volunteers...
Generated Summary	– Madonna’s new film, W.E., about King Edward VIII and his affair with Wallis Simpson, got a standing ovation at its premiere at the Toronto International Film Festival last night, and the Queen of Pop didn’t mince words when talking about how she felt about the “very punk rock” King Edward. “I thought he was quite rebellious and cutting edge in his point of view about life and about how to run the empire and using the Sex Pistols was a perfect marriage,” she said, according to the Guardian. The song “God Save the Queen” from the punk band was on the soundtrack for the film, Reuters notes, and Madonna also said she could sympathize with the late King Edward because he was an outsider, an American living in London. "I empathize with Wall is. Public figures or icons are often just reduced to a sound-bite, just a handful of attributes. I think that’s why I feel so strongly about Wallis," she said. The film, which debuted at the Venice Film Festival earlier this month, has received some negative reviews, but Madonna says she doesn’t mind, so long as they’re directed at the film and not at her, the Guardian reports. " I can tell when people are reviewing my film and when they’re reviewing me personally. I welcome criticismsof my film when it’s viewed as an artistic form and not when people were mentioning things about my personal life or my achievements in any other field because they’re irrelevant to the film. So when they stick to the movie then I do care, I pay attention to it.”
Extracted Evidence	"I welcome criticisms of my film when it’s viewed as an artistic form and not when people are mentioning things about my personal life or my achievements in any other field because they’re irrelevant to the film. (0.3320)
	Madonna said: "I thought he (King Edward VIII) was quite rebellious and cutting edge in his point of view about life and about how to run the empire and using the Sex Pistols was a perfect marriage." (0.2817)
	She had a wonderful time at the festival and was especially delighted that she got to spend so much time with her fans in front of the theatre which is a famous tradition at the Festival ... We are still trying to figure out who and why anyone would ask the volunteers to turn away from Madonna. (0.1724)
	"I can tell when people are reviewing my film and when theyŕe reviewing me personally," Madonna said when asked whether she cared about what critics thought. (0.1604)
	In Venice, where the film premiere earlier this month, the Queen of Pop said there were "elements of myself" in the film, and said she could sympathise with Wallis as an outsider, an American living in London. (0.1387)
	"I can tell when people are reviewing my film and when they’re reviewing me personally," she said. (0.1290)
	Madonna described King Edward VIII as "very punk rock," at the North American opening of her film W.E about his infamous romance with American divorcee Wallis Simpson - and his subsequent abdication. (0.1199)
	Speaking at the Toronto Film Festival, the singer said she wanted reviewers to focus on her work as a director – not on her personal life. (0.1144)
	The singer-turned-film-director was explaining how an anarchist anthem by punk band The Sex Pistols ended up on the soundtrack for her second directorial work, starring British actors James D’Arcy and Andrea Riseborough. (0.1059)
	TORONTO (Reuters) - Madonna on Monday said she said she doesn’t mind any criticism of her filmmaking abilities, so long as it is directed at her movie and not at herself. (0.1018)

Table 9: Examples on the MULTI-NEWS dataset (2).

Gold Summary	– A British man who hacked his way into Pentagon computers looking for information on UFOs has lost his final appeal against extradition. Gary McKinnon, who suffers from Asperger’s syndrome, is now expected to be sent to the US for trial within weeks, the Independent reports. The 43-year-old Londoner could face up to 60 years in jail if found guilty of what US authorities call the biggest military hack in history. Opposition lawmakers and McKinnon’s supporters slammed the decision to reject the appeal, citing medical reports that said deportation was likely to make him suicidal. "To force a peaceful, vulnerable, misguided UFO fanatic like Gary thousands of miles away from his much-needed support network is barbaric," his mother told the Telegraph. "This is a cruel and miserable decision.
Generated Summary	– Gary McKinnon, the UK hacker accused of hacking into NASA and the military to find evidence of UFOs, is now "at an even higher risk of self- harm and suicide" and will kill himself in order to be extradited to the US, according to a new medical report. McKinnon has Asperger’s syndrome and has a “ fixed-psychological conviction he will kill myself in preference to being extradited,” the report says, adding that suicide is now an “almost certain inevitability.” The British government "is terrified of speaking up to America, and now they are allowing vulnerable people to be pursued for non-violent crime when they should be going after terrorists," McKinnon’s mother told the Telegraph. "This is a cruel and miserable decision. to force a peaceful, vulnerable, misguided UFO fanatic like Gary thousands of miles away from his much-needed support network is barbaric. If the severity of Gary’s medical condition isn’t sufficient to prevent his extradition, I can’t imagine what is." McKinnon could face up to 60 years in prison in the US.”
Extracted Evidence	She added: "This government is terrified of speaking up to America, and now they are allowing vulnerable people to be pursued for non-violent crime when they should be going after terrorists. (0.2463)
	If the severity of Gary’s medical condition isn’t sufficient to prevent his extradition, I can’t imagine what is." (0.1838)
	The medical report said Mr McKinnon, who is known to suffer Asperger’s syndrome, a form of autism, also has a “fixed-psychological conviction he will kill himself in preference to being extradited”. (0.1696)
	Mrs Sharp said: "To force a peaceful, vulnerable, misguided UFO fanatic like Gary thousands of miles away from his much-needed support network is barbaric. (0.1637)
	But following the failure of Mr McKinnon’s bid to have his case heard at the Supreme Court, he concluded that he was now "at an even higher risk of self harm and suicide", something he said was a "real probability". (0.1603)
	Extracts of Prof Turk’s report, disclosed in the Daily Mail, show that he recorded in September that Mr McKinnon, 43, suffered from a "’very serious Major Depressive Disorder.... aggravated and complicated by anxiety and panic attacks with multiple psychosomatic symptoms on a background of his having Asperger’s syndrome". (0.1517)
	The family said last night that Mr McKinnon, who could be sentenced to up to 60 years in prison in the US, was "at risk of suicide" after being told there will be no 11th hour reprieve. (0.1470)
	To force a peaceful, vulnerable, misguided UFO fanatic like Gary thousands of miles away from his much-needed support network is barbaric. (0.1392)
	In July Mr McKinnon went to the High Court in an attempt to get the extradition order overturned but was told being sent for trial in the US was "a lawful and proportionate response" to his actions. (0.1350)
	His mother, Janis Sharp, was "extremely worried" about her son’s mental state and said the Government and Mr Johnson should "hang their heads in shame" for caving in to American pressure. (0.1307)

Table 10: Examples on the MULTI-NEWS dataset (3).