Exploring Explainable Selection to Control Abstractive Summarization

Haonan Wang[1], Yang Gao[1,*], Yu Bai[1], Mirella Lapata[2], Heyan Huang[1,3]

[1]School of Computer Science and Technology, Beijing Institute of Technology
[2]Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh
[3] Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
{hnwang, gyang, yubai}@bit.edu.cn, mlap@inf.ed.ac.uk, hhy63@bit.edu.cn

Abstract

Like humans, document summarization models can interpret a document's contents in a number of ways. Unfortunately, the neural models of today are largely black boxes that provide little explanation of how or why they generated a summary in the way they did. Therefore, to begin prying open the black box and to inject a level of control into the substance of the final summary, we developed a novel select-and-generate framework that focuses on explainability. By revealing the latent centrality and interactions between sentences, along with scores for sentence novelty and relevance, users are given a window into the choices a model is making and an opportunity to guide those choices in a more desirable direction. A novel pair-wise matrix captures the sentence interactions, centrality and attribute scores, and a mask with tunable attribute thresholds allows the user to control which sentences are likely to be included in the extraction. A sentence-deployed attention mechanism in the abstractor ensures the final summary emphasizes the desired content. Additionally, the encoder is adaptable, supporting both Transformer- and BERT-based configurations. In a series of experiments assessed with ROUGE metrics and two human evaluations, ESCA outperformed eight state-of-the-art models on the CNN/DailyMail and NYT50 benchmark datasets.

Introduction

The ability to generate summaries of documents is a valuable tool over the past several years, and neural networks have been responsible for a step-change in the quality of both extractive and abstractive summarization. Extractive methods simply draw out and concatenate the key topic sentences in a document (Nallapati, Zhai, and Zhou 2017; Zheng and Lapata 2019), while abstractive techniques reorder words and sentences and even generate new language to, hopefully, produce a concise and eloquent piece of the given content (See, Liu, and Manning 2017; Celikyilmaz et al. 2018; Wang et al. 2019). However, despite recent advancements, modelling concepts that span more than a few sentences, i.e., long-range contexts, still a challenging task. Moreover, current models provide little to no explanation of the interpretation they took away from parsing a document and why they chose to summarize its content in the way that they did.

*Corresponding Author.

Currently, two broad strategies for tackling this problem are explored. The first is to use pre-trained language model, such as ELMo (Peters et al. 2018), OpenAI GPT (Radford et al. 2018) and BERT (Devlin et al. 2018), have achieved state-of-the-art performance on long-range contextual learning and various NLP tasks, such as QA (Xu et al. 2019) and summarization (Liu and Lapata 2019; Zhang, Wei, and Zhou 2019). The other idea is to use a *select and generate* framework, where an extractor selects salient sentences, then an abstractor generates a summary. The most recent frameworks based on this hybrid paradigm either follow a two-stage pipeline (Chen and Bansal 2018; Sharma et al. 2019) or an end-to-end learning approach (Hsu et al. 2018; Shen et al. 2019; Gehrmann, Deng, and Rush 2018). The most appealing advantage is to explicitly obtain desirable content of the sources, such as entity-aware selection (Sharma et al. 2019) or word selection through latent switch variables (Gehrmann, Deng, and Rush 2018; Shen et al. 2019).

These approaches perform abstractive summaries which largely rely on selecting informative content (extractor) as well as aggregating into a summary in line with linguistic expression (abstractor). But, currently, the extractors are largely black-box decisions without a rationale of what is informative content. Peyrard (2019) proposed rigorous definitions of the concepts in summarization, including *redundancy*, *relevance* and *informativeness*. While, more in-depth investigation of these concepts are needed for them to be truly useful to document summarization. For instance, we need to better understand inter-relations between sentences with respect to these attributes. We need methods for identifying the sentence *informativeness*, identifying whether a sentence is *relevant* to a document and, if so, to what extent. Another importance influence is the *novelty* of the contribution a sentence makes to a summary.

Moreover, abstractive summarization suffers from a major problem known as hallucination, where the model generates fictional content (Maynez et al. 2020). The cause is believed to due to misrepresenting content in a batch of input documents and fusing concepts across those documents when generating abstractive summaries. Additionally, some of the new terms introduced are thought to come from background knowledge, not from the current inputs. Some researchers have attempted to alleviate this problem with pointer mechanisms to desired content (See, Liu, and Manning 2017; Wang

et al. 2019; Celikyilmaz et al. 2018) or by interpolating nearest neighbors computed from the inputs (Khandelwal et al. 2019) and so on. However, due to the scattered tracts of information in long documents, it is inevitable that some irrelevant and unnecessary content will be picked up when generating summaries. As a result, there is potential for an abstractive summary to completely depart from the gold summary into a fictional hallucination and, unfortunately, this is difficult to control.

Therefore, to reveal more of the inner workings of these black-box models so as to inject a level of control into the substance and integrity of the final summary, we developed a novel select-and-generate framework, called **ESCA** (Loosely, means Explainable Selection module to Control the generation of Abstractive summaries), that focuses on explainability. The key to the framework is an interaction matrix that highlights the decisions made about each sentence, which can be decoupled into three explicit components, the informativeness of a sentence, its relevance to the substance of the document, and its novelty with respect to the accumulated summary representation. A novel pair-wise ranking extractor then selects sentences for extraction, favoring the complex relations within each sentence pair and its potential influence over the summary. To avoid hallucinations, a sentence-deployed attention mechanism in the abstractor, but populated with values from the extractor, ensures the abstractive summary focuses on both correct and desired concepts. As such, the extractor and abstractor are seamlessly integrated with a deployment-based pointer in an end-to-end manner. Further, which content is selected for extraction can be controlled by setting thresholds for novelty and relevance and applying a mask that adjusts the probability of extraction accordingly.

In summary, our contributions include: 1) an explainable content selection module for document summarization; 2) the ability to extract the appropriate content for generating a desired summary based on explicit and quantified measures of informativeness, novelty and relevance to the final summary; 3) automatically creating synthetic datasets w.r.t novelty and relevance for exercising controllable inference without the need to retrain the entire system. A series of experiments assessed with ROUGE metrics and two human evaluations demonstrate that ESCA provides summaries of higher quality than eight state-of-the-art models on the CNN/DailyMail and NYT50 benchmark datasets.

Related Work

As opposed to extractive summarization, where all but the most salient and meaningful sentences are removed to reduce an entire document down to a short summary of its contents, abstractive summarization generating new or rephrased words and sentences to produce the summary. Headline generation is a subtask of abstractive summarization and, in this area, seq2seq models have largely accomplished the goal of generating snappy and expressive headlines. (Nallapati et al. 2016; Zhou et al. 2017; Shen et al. 2019; Wang et al. 2019). However, summarizing content with notions that span more than a few sentences, i.e., long-range contexts, with abstractive techniques remains a significant challenge.

Models called pointer-generator overcome this explainability problem to some extent by using attention as a pointer, conditioned on contextual information, jointly determine which language to select/generate based on probability (Vaswani et al. 2017; See, Liu, and Manning 2017). Further, pointer generators can operate at either the word level (Wang et al. 2019; Çelikyilmaz et al. 2018) or the sentence level (Chen and Bansal 2018; Sharma et al. 2019). At the word level, Zhou et al. (2017) used soft gating on the source document to produce summaries, while Gehrmann, Deng, and Rush (2018) pre-trained a sequential word selector to constrain attention from the source document. Hsu et al. (2018) updated word attention by considering importance at the sentence level. Among the sentence-level models, Tan, Wan, and Xiao (2017) used a graph-based attention mechanism with an abstractive model leveraged by improving salient sentence selection. Li et al. (2018) achieved the same goal with an information selection layer consisting of global filtering and sentence selection modules. You et al. (2019) subsequently improved salience attention by introducing a Gaussian focal bias to better inform the selection process.

In terms of the generation process, a single text can be summarised in diverse target sequences with different focus (Cho, Seo, and Hajishirzi 2019). To tackle this issue, Shen et al. (2019) used decoupled content selection to allow fine-grained control over the generation process. In our framework, we leverage the benefits of a pointer generator model for selection, but we also explore explaining the content to be selected for extracted as a way to control the generation process so as to produce a desirable summary.

Background: Encoder-Decoder Framework

In summarization, consider a sequential input $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_n\}$ of n number of words with j as the index of the input, the shortened output, i.e., the summary, is denoted as $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_m\}$ with m number of words, where t indicates the position of the output. The basic structure is based on Transformer unit composed of a stack of N identical layers, and each layer has two sub-layers: the first is a self-attention sub-layer $\boldsymbol{h}_1^l$, and the second is a feed-forward sub-layer $\boldsymbol{h}_2^l$ with a depth of l. Then, a multi-head operation follows the feed-forward sub-layer. The final output sequence of the encoder is denoted as $\boldsymbol{Z}_e$. The decoder consists of a similar stack of N identical layers. But, in addition to the two sub-layers for each layer, the decoder includes a third sub-layer, that performs multi-head attention over the output of the encoder stack. Following this procedure, an attention value between the decoder position vector $\boldsymbol{s}_t$ and the encoder sequence output $\boldsymbol{Z}_e$ is calculated for each source position. Attention on the source input at the decoder position t is then calculated with the formula $\boldsymbol{\alpha}_t = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_m}})$, where $\boldsymbol{Q}$ is $\boldsymbol{s}_t$ and $\boldsymbol{K}$ is $\boldsymbol{Z}_e$. The context vector at the decoding position t is $\boldsymbol{h}_t^* = \boldsymbol{\alpha}_t \boldsymbol{Z}_e$. From here, the decoder generates a summary, called the target summary, from a vocabulary distribution $P_{vocab}(w)$ through the following process:

$$\begin{aligned} P_{vocab}(w) &= P(y_t | \boldsymbol{y}_{<t}, \boldsymbol{x}; \theta) \\ &= \text{softmax}(\boldsymbol{W}_2(\boldsymbol{W}_1[\boldsymbol{s}_t, \boldsymbol{h}_t^*] + \boldsymbol{b}_1) + \boldsymbol{b}_2) \end{aligned} \quad (1)$$

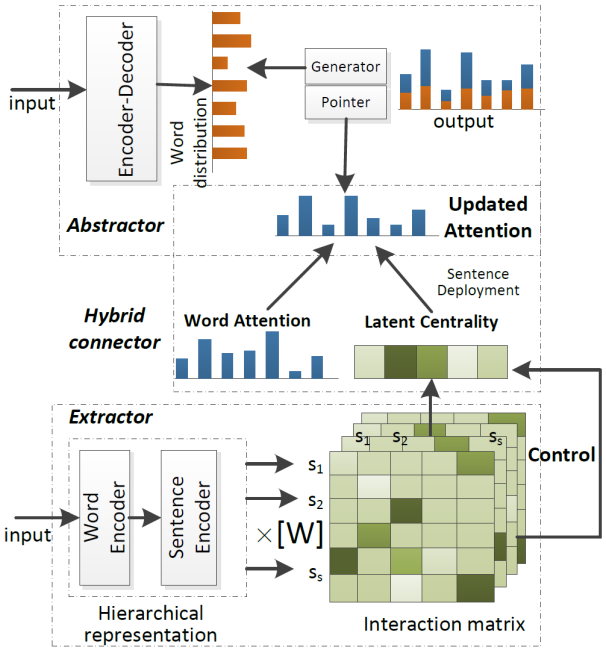

Figure 1: The framework of the proposed method

The Proposed Model

Our model is an end-to-end hybrid summarization framework. The architecture and summarization process is demonstrated in Figure 1. It comprises: (1) a pair-wise extractor that incorporates a sentence interaction matrix and uses latent centrality; (2) the abstract generation is guided by the sentence deployed attention as a pointer to hybrid with a Pointer Generator (PG) abstractor and (3) controllable and tunable interaction matrix that explains selecting different content that will affect the final abstractive summary.

The Extractor

The encoder of our framework can be equipped with any vector-based neural networks. In this paper, we implement Transformer and BERT, showing its flexibility. The detailed settings are depicted in the experimental section.

Explaining the Interaction Matrix As stated by Nallapati, Zhai, and Zhou (2017), complex relations exist in each sentence pair that add informativeness, novelty and relevance to the content of a document. Inspired by the relations, our sentence interaction matrix, $\boldsymbol{Q}^s$, reflects these complexities along with the similarity of the sentence pair. In $\boldsymbol{Q}^s$, s denotes the number of sentences. We further note that "interactions" have a direction, because, as suggested by Zheng and Lapata (2019), the contribution induced by two sentences' relations to their respective importance as a summary can be unequal. Furthermore, the mutual influence of the sentence pair may have a different direction, which is grounded in the theory of discourse structure (Mann and Thompson 1988). The directional influence of sentence j to sentence i is denoted as q_{ij} in the interaction matrix $\boldsymbol{Q}^s$, including informativeness of the sentence s_i, relevance of s_i to the document and novelty of s_i to existing summary.

Specifically, the *informativeness* refers to how important and informative of a specific sentence i, the *relevance* indicates to what extent the sentence i is relevant to a document d, and the *novelty* means what the new information of the sentence i contributes to the summary. Therefore, the above attributes explicitly decouple the interaction matrix, showing explainable meanings to the summary, and the directional influence score q_{ij} in $\boldsymbol{Q}^s$ is:

$$q_{ij}(\boldsymbol{h}_i, \boldsymbol{s}_i, \boldsymbol{h}_j, \boldsymbol{d}) = \sigma\Big(\underbrace{\boldsymbol{W}_c \boldsymbol{h}_i}_{\text{informativeness}} + \underbrace{\boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{W}_r \boldsymbol{d}}_{\text{relevance}} + \underbrace{\boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{W}_s \boldsymbol{h}_j - \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{W}_n \tanh(\mathbf{a}_i)}_{\text{novelty}} + \boldsymbol{b}_{\text{matrix}} \Big) \tag{2}$$

where $\boldsymbol{h}_i$ is the representation of sentence i, and $\boldsymbol{d}$ is the vector of the input document. $\mathbf{a}_i$ is the accumulated summary representation w.r.t the current sentence i and is $\mathbf{a}_i = \frac{1}{s} \sum_{t=1}^{i-1} \sum_{k=1}^{s} \boldsymbol{h}_t \times q_{tk}$, where q_{tk} represents the influence of sentence k to sentence t. Note that the novelty roughly decreases with the latter sentences as normal summary positioned in the front position. σ is a sigmoid function, and $\boldsymbol{W}_c, \boldsymbol{W}_s, \boldsymbol{W}_n, \boldsymbol{W}_r, \boldsymbol{b}_{\text{matrix}}$ are trainable parameters.

Latent Centrality Calculation The interaction matrix $\boldsymbol{Q}^s$ stores the mutual influence of each sentence pair, which helps to estimate the overall importance of the sentence. There are several summarization models for computing the centrality of a sentence, including the graph-based TextRank (Mihalcea and Tarau 2004) and LexRank. Additionally, Tan, Wan, and Xiao (2017) drew on a similar idea with a model that determines sentence salience via graph-based attention. In our end-to-end setting, the interaction matrix $\boldsymbol{Q}^s$ is directly transformed into the sentences' distribution, which is then converted into a centrality vector:

$$\boldsymbol{c} = \boldsymbol{Q}^s \boldsymbol{W}_q \tag{3}$$

where $\boldsymbol{c} = [c_1, \cdots, c_s] \in \mathcal{R}^s$ is the sentence centrality, and $\boldsymbol{W}_q \in \mathcal{R}^{s \times 1}$. In our experiments, we truncated the sentence number s in the documents to a maximum of 50.

Pair-wise Learning Extractor The process of extraction can be framed as a classification problem. Nallapati, Zhai, and Zhou (2017); Hsu et al. (2018); Liu and Lapata (2019) all use a point-wise ranking approach in which sentences are encoded as hidden representations. Then, a binary classifier is trained on those representations to predict whether or not they are suitable for the summary. However, because point-wise learning is not yet powerful enough to accurately reflect the interactions between sentences, we introduced a new pair-wise loss function supported by inter-sentence labels that helps the extractor decide the summary classification. More specifically, first, each sentence is labeled as described in Appendix A[1]. Then, the inter-sentence label for each sentence

[1]Appendices are included in https://arxiv.org/pdf/2004.11779.pdf

pair $\hat{P}_{ij}$ is marked with $\{0, 1\}$, where 1 indicates the sentence i has been selected for the summary, but sentence j has not; 0 indicates the opposite – that sentence j has been selected for the summary while i has not. To adapt our supervised system to summarization, the predicted co-occurrence probability r_{ij} of sentence i needs to be calculated. The formula is $\sigma(c_i - c_j)$, and the loss function is then defined as

$$\mathcal{L}_{\text{ext}} = -\sum_{i=1}^{m}(\hat{P}_{ij}\log r_{ij} + (1-\hat{P}_{ij})\log(1-r_{ij})) \quad (4)$$

The Abstractor

The abstractor is based on a pointer-generator network containing two sub-modules: the pointer network and the generation network. These two sub-modules jointly determine the probability that a word will be included in the final generated summary. Our proposed model essentially leverages this configuration that integrates a new sentence deployed pointer, introducing the selected content flow into the generation network in the hybrid framework.

Sentence Deployed Pointer Generator The pointer network uses attention as a pointer to select segments of the input as outputs (Vinyals, Fortunato, and Jaitly 2015). As such, a pointer network is a suitable mechanism for extracting salient information, while remaining flexible enough to interface with a language model for generating an abstractive summary (See, Liu, and Manning 2017).

In our pointer network, the selected segments of input can be updated by the extractor with respect to their extractive-oriented centrality of each sentence. To influence the sequence generation, sentence importance needs to be deployed to the word level. The deployment should determine how much information flow is delivered to the word-level generation, at the same time considering the importance of the derived sentence. With these values, the pointer can seamlessly link the extractor with the abstractor via the hybrid connector.

The pointer is taken by the attention distribution that will be updated by our proposed *hybrid connector*. The hybrid is achieved by the sentence deployment attention mechanism, which controls the generation process by focusing on what the selected content explicitly conveys. The equation for calculating the pointer distribution leveraged by sentence deployed attention is as follows,

$$\hat{\alpha_t^n} = \frac{\alpha_t^n(1+p_{sen}c_{m_n})}{\sum \alpha_t^n(1+p_{sen}c_{m_n})} \quad (5)$$
$$p_{sen} = \sigma(\boldsymbol{W}_{sel}\boldsymbol{E}_{sel}^t + \boldsymbol{b}_{sen})$$

where c_{m_n} denotes the score of the sentence m that the word n belongs to. $\boldsymbol{E}_{sel}^t$ is the representation of the selected sentence m at the decoding step t. p_{sen} decides the degree of influence a sentence will have on the summary. $\boldsymbol{W}_{sel}$ is a trainable parameter. Additionally, the generation probability p_{gen} is modified with

$$p_{gen} = \sigma(\boldsymbol{W}_{h^*}\boldsymbol{h}_t^* + \boldsymbol{W}_s\boldsymbol{s}_t + \boldsymbol{b}_{gen}) \quad (6)$$

The pointer is taken based on the updated attention distribution $\hat{\alpha_t}$ over the source text, and the final output distribution is combined as follows:

$$P_{\text{final}}(w) = p_{gen}P_{vocab}(w) + (1-p_{gen})(\sum\nolimits_{j:w_j=w}\hat{\alpha_{t,j}}) \quad (7)$$

The basic generator objective is derived by maximizing the likelihood of pointer-generator during training, given a reference summary $y^* = \{y_1^*, y_2^*, \cdots, y_{m'}^*\}$ for a document x, and the training objective is to minimize the negative log-likelihood of the target word sequence:

$$\mathcal{L}_{abs} = -\sum_{t=1}^{m'}\log P_{\text{final}}(y_t^*|y_1^*, \cdots, y_{t-1}^*, \boldsymbol{x})$$

Overall, the learning objective is $\mathcal{L} = \mathcal{L}_{ext} + \mathcal{L}_{\text{abs}}$.

Controllable Inference

Since the interaction matrices capture the inter-sentence relations leveraged by different explainable aspects, as outlined in the section of extractor, the overall centrality has the potential to reflect the aspects of the content and, in turn, explain how selecting one sentence over another will influence the final abstractive summary. Hence, to explore this explainability and fine-tune which sentences to extract to produce the most desirable abstractive summary, the sentence selections can be manipulated through several mask matrices $\boldsymbol{M}$ based on controllable thresholds for novelty ϵ_n and relevance ϵ_r versus each sentence's scores in these areas. Note that the informativeness is not proper to be controlled since it only relates to the sentence itself without any interaction with other sentences or the document. The fine-tuning control is applied using the following equation:

$$\hat{\boldsymbol{Q}}^s = \boldsymbol{Q}^s \odot \boldsymbol{M}, \text{ where } M_{ij} = \begin{cases} 1, val \geq \epsilon \\ 0, val < \epsilon \end{cases} \quad (8)$$

where $\odot$ is element-wise multiplication, and val is the $\sigma(novelty)$ or $\sigma(relevance)$ calculated from Eq.(2). In this way, the mask matrices, $\boldsymbol{M}_n$ (novelty) or $\boldsymbol{M}_r$ (relevance), can be adjusted to control which content to focus on. Once satisfied with the sentence selections, the document graph is then reshaped to align with the different mask matrices, and the summary selection is changed because of the revised centrality. Generating the d ifferent summaries for the final output based on the controllable masks is done without additional training. In turn, enforcing sentence-deployed attention through Eq.(5) tells the abstractor what to focus on as it infers and generates the abstractive summary.

Experiments

In this section, we describe the datasets used in the experiments, our setup, implementation details[2] and evaluation methods, and analyze the result.

Datasets We evaluated our models and baselines on two benchmark datasets, namely the CNN/DailyMail news set (Hermann et al. 2015), and the New York Annotated Corpus

[2]Our code and dataset samples are available on https://github.com/Wanghn95/Esca_Code

(NYT) (Sandhaus 2008). The CNN/DailyMail dataset[3] contains news articles and associated highlights as summaries. We followed the standard splits 90,266/1220/1093 for the training, validation and testing sets for the CNN dataset and 196,961/12,148/10,397 for the DailyMail dataset. We did not anonymize the entities, and the datasets were pre-processed following See, Liu, and Manning (2017). The NYT dataset[4] contains 110,540 articles with abstractive summaries, which were divided into 100,834 articles for the training set and 9,706 for the test set, following Durrett, Berg-Kirkpatrick, and Klein (2016). We also filtered the raw datasets by eliminating the documents with summaries shorter than 50 words. The filtered test set, called NTY50, included 3,421 examples. The abstractor processed the input by truncating the source documents to 400 tokens for CNN/DailyMail and 800 for NYT. As discussed in Liu and Lapata (2019), the NYT test set contains longer and more elaborate summaries than the CNN/DailyMail set, whose summaries are largely extractive and mostly concentrate on the beginning of the documents. All sentences were split with the Stanford CoreNLP toolkit (Manning et al. 2014). We used ROUGE as the evaluation metric (Lin 2004)[5], which measures the quality of a summary by computing the overlapping lexical elements between the candidate summary and a reference summary. Following previous practice, we assessed R-1 (unigram), R-2 (bigram) and R-L (longest common subsequence).

ESCA-Transformer was trained with a 6-layer transformer with 8 heads. The hidden size was set to 512, and the feed-forward dimension for the multi-head attention was set to 1024. We used dropout with a probability of 0.2 prior to the linear layers. The learning rate for the pointer-generator was 0.15 with a batch size for the encoder of 32 and a beam size for the decoder of 4. The learning rate of both of them was 0.15. At the testing phase, we limited the length of the summary to 120 words. The model was trained with an early stopping and length penalty imposed on the validation set.

ESCA-BERT followed the settings specified by Liu and Lapata (2019). Specifically, we inserted `[CLS]` tokens at the start of each sentence, and also used two-interval segment embeddings $[E_A]$ or $[E_B]$ to distinguish between multiple sentences in a document. The `[CLS]` then learned the sentence embedding. Position embeddings in the BERT model had a 512 length limit. We used the standard 'BERT-base-uncased' version of BERT [6]. Both the source and target tokens were tokenized with BERT's subwords. The hidden size of the transformer layers was 768, and all the feed-forward layers had 2048 hidden units. One transformer layer in the extractor with 8 heads and a dropout of 0.1 was dedicated to producing the sentence representations. We used the trigram block trick (Paulus, Xiong, and Socher 2017) to prevent duplicates. The abstractor was trained over 15k iterations for the NYT dataset and 100k iterations for CNN/DM with label smoothing loss (Szegedy et al. 2016) at a factor of 0.1. Moreover, dropout with a probability of 0.2 was applied prior to the linear layers. The decoder contained 6 transformer layers. We used separate learning rates of 0.002 and 0.2 for the BERT encoder and Transformer decoder, respectively. The settings for the decoding process were the same as those outlined for the Transformer-based model above. The final model contained 180M parameters.

Comparative Models Each of the following state-of-the-art models follow the "select and generate" style (BERTSUMabs excluded), as Section outlined in related work. The eight chosen comparators were: PG+Coverage is a BiGRU-based seq2seq model integrated with a pointer network and an additional coverage mechanism (See, Liu, and Manning 2017). Select-Reinforce (Chen and Bansal 2018), which reinforces the extraction of important sentences with a reward function based on a summary rewrite evaluation metric. Inconsistency-Loss (Hsu et al. 2018) includes a loss function that uses sentence-level attention to modulate word-level attention for generating summaries. Bottom-Up (Gehrmann, Deng, and Rush 2018) uses an extractive encoder as a content selector to constrain word attention for the abstractive summarization. ExplicitSelection (Li et al. 2018) is an extended version of the vanilla seq2seq model with a soft information selection layer to control information flow. SENECA (Sharma et al. 2019) selects entity-aware sentences and then connects them abstract generation based on reinforcement learning. BERTSUMabs and BERTSUMextabs are developed by Liu and Lapata (2019), which are not "select-and-generate models". BERTSUMextabs adopts a two-stage fine-tune of extractor and abstractor.

Quantitative Analysis

The overall results are presented in Table 1 and 2. We observed that ESCA-Transformer had competitive performance to most of the baselines, and ESCA-BERT outperformed all the strong state-of-the-arts on both datasets in all metrics. Relatively speaking, ESCA-BERT has a higher improvement (1.20% comparing with the most advanced BERTExtAbs) in R-2 metric on the NYT dataset whose gold summaries are longer and more abstractive than the ones in the CNN/DailyMail datasets (Liu and Lapata 2019). It indicates that the ESCA model has advantage on generating long-length and fluent summaries.

Ablation Studies of the Extractor To investigate the effectiveness of our pair-wise ranking strategy, we compared it with its counterpart - point-wise ranking. The probability of a sentence as a extractive summary was calculated by $\sigma(c_i)$ where c_i is derived from Eq. (3). The point-wise ranking loss was then computed through the cross entropy of the predicted score with the gold label. The upper block of Table 3 focuses on the extraction performance of the CNN/DailyMail dataset. It is clear that our model with pair-wise ranking is largely superior to the point-wise extractor in terms of the ROUGE scores, with relative improvements ranging from 11.0% to 11.55%. Since our study focuses more on the global sentence distributions over the documents, we selected the top 6 sentences for a more granular analysis of the results. Our observations suggest that the overall distribution of the pair-wise extractor is more likely to be close to the gold summary.

[3]https://cs.nyu.edu/~kcho/DMQA/

[4]https://catalog.ldc.upenn.edu/LDC2008T19

[5]Implemented by pyrouge package based on ROUGE1.5.5.

[6]https://github.com/google-research/bert

Models	R-1	R-2	R-L
PG+Coverage	39.53 *	17.28*	36.38*
Select-Reinforce	40.88 *	17.80*	38.54*
Inconsistency-Loss	40.68*	17.97*	37.13*
Bottom-Up	41.22*	18.68*	38.34*
Explicit-Select †	41.54	18.18	36.47
SENECA †	41.52	18.36	38.09
BERTSUMabs	41.72*	19.39*	38.76*
Ours			
ESCA-Transformer	41.34*	18.50*	37.94*
ESCA-BERT	**42.01**	**19.52**	**39.07**

Table 1: The ROUGE scores for the abstractive summaries from the CNN/DailyMail datasets. Results marked with a † mark are taken from the their corresponding papers. $*$ indicates a significant difference between the comparing model and the ESCA-BERT (at $p < 0.1$, using a pairwise t-test).

Models	R-1	R-2	R-L
PG+Coverage	43.71*	26.40 *	37.79*
Bottom-Up	47.38 *	31.23*	41.81*
SENECA †	47.94	31.77	44.34
BERTSUMabs	48.92*	30.84*	45.41*
BERTSUMextabs	49.02*	31.02*	45.55*
Ours			
ESCA-Transformer	47.63*	30.10*	43.94*
ESCA-BERT	**49.41**	**32.22**	**45.83**

Table 2: The ROUGE scores for the abstractive summaries from the NYT50. Results marked with a † are taken from their corresponding papers. $*$ indicates a significant difference between the comparing model and the ESCA-BERT ($p < 0.1$, using a pairwise t-test).

This verifies that the pair-wise ranking has a noticeable effect on the quality of the extracted summary.

According to Vaswani et al. (2017), inter-sentence relations can also be captured by multiple stacked self-attention layers. Therefore, we replaced the interaction matrix with a 2-layer self-attention mechanism to build a counterpart variant of ESCA's extractor, called $\text{Extractor}_{\text{self-attention}}$. We selected the top 3 sentences for evaluation and, report the comparative results in the lower block of Table 3. Based on the scores, there was no significant difference in the quality of the selection. However, self-attention may not be able to adequately explain why each sentence was selected.

Controllability

Synthetic Datasets To evaluate the impact of attending to relevance and novelty, we created two sample datasets based on the testing set of the two original CNN/DailyMail sets. The dataset for relevance was constructed by adding a title as part of the original gold summaries to increase the relevance between the summaries and the input document. In terms of novelty Zhou et al. (2018) found that the CNN/DailyMail gold summaries favor leading sentences, which may not cover

Models	R-1	R-2	R-L
$\text{Extractor}_{\text{PointWise}}$	32.68	15.41	30.33
$\text{Extractor}_{\text{PairWise}}$	36.41	17.19	33.68
$\text{Extractor}_{\text{self-attention}}$	42.8	20.1	39.2
$\text{Extractor}_{\text{interaction-matrix}}$	42.7	20.0	39.2

Table 3: ROUGE scores from CNN/DailyMail for the extractive summaries of ESCA-BERT model and its counterparts.

Control	Threshold	R-1	R-2	R-L
Novelty	$\epsilon_n = 0$	44.78	35.39	42.25
	$\epsilon_n = 0.3$	45.66 ↑	36.28 ↑	43.05 ↑
	$\epsilon_n = 0.4$	45.26 ↑	36.08 ↑	42.67 ↑
	$\epsilon_n = 0.5$	45.28 ↑	35.90 ↑	42.71 ↑
Relevance	$\epsilon_r = 0$	41.35	18.50	38.57
	$\epsilon_r = 0.3$	41.41 ↑	18.57 ↑	38.62 ↑
	$\epsilon_r = 0.5$	41.52 ↑	18.67 ↑	38.55 ↓
	$\epsilon_r = 0.7$	41.27 ↓	18.44 ↓	38.43 ↓

Table 4: Controllability: the ROUGE scores from the CNN/DailyMail datasets for different thresholds of novelty ϵ_n and relevance ϵ_r (absolute decrease/increase performance over $\epsilon_n = \epsilon_r = 0$ is denoted by ↑/↓).

the content of the document comprehensively. Hence, we employed an advanced unsupervised extractive summarization method called, PacSum (Zheng and Lapata 2019), to discover more diverse summaries. PacSum disregards the first five sentences in an article and then selects the top-3-ranked sentences from the remainder of the input document for extraction. Then, the original gold summaries are complemented with the novel content. To explore the explainable selection w.r.t relevance and novelty as mentioned in Eq.(2), we manually set different thresholds to construct the masking matrices with Eq.(8). We used ROUGE F1 to evaluate the influence of these controllable thresholds on the two synthetic datasets. The scores are shown in Table 4. The results illustrate the control over different scales of novelty is indeed able to generate diverse summaries, while a relevance score of $\epsilon_r = 0.5$ (except for R-L) generated the best summaries. However, there is always a trade-off between controllability and summary quality.

It also shows that the ROUGE scores are varied weakly because of two reasons. First, the controlled summaries must preserve the informative content that the original ESCA has. The ROUGE score cannot be largely changed, otherwise, the summary can be wrong. Second, ROUGE score has drawbacks for evaluation w.r.t the overlapped vocabularies. To further verify the effectiveness of these two controllable parameters, we conducted a human evaluation with novelty and relevance criteria in the following section and some examples are provided in Appendix C for further inspection.

Explainable Matrix over Control To explicitly demonstrate the power of the interaction matrix $\mathbf{Q}_s$ how the influence of novelty and relevance explains the final abstractive summaries, we visualize them as heatmaps, shown in Figure 2. From these, we find that novelty can move the focused centrality from leading sentences to scattered spans of the

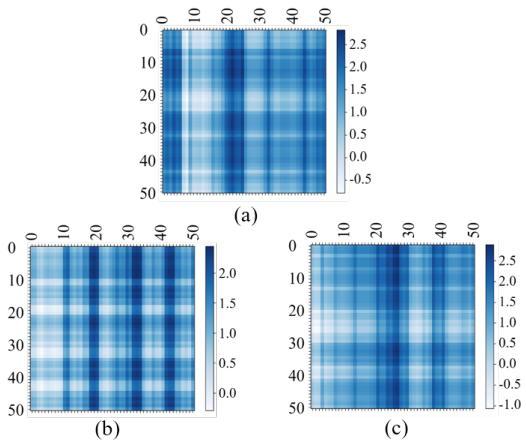

Figure 2: (a) A visualization of the interaction matrix Q_s. (b) The reshaped matrix controlled by novelty. (c) the reshaped matrix according to relevance. For simplicity, the thresholds were set to $\epsilon_n = \epsilon_r = 0.5$.

document which capture novel content (Figure 2(b)). Relevance slightly decreases the effect of leading sentences, while enhancing the centrality of salient content (Figure 2(c)).

Human Evaluation

Two separate human evaluations were conducted. The first was a question answering (QA) test and the second was to assess the quality of the summaries. Both types of evaluations were conducted on the Amazon Mechanical Turk platform.

QA evaluation have been previously used to evaluate the quality of a summary and document compression (Clarke and Lapata 2010; Narayan, Cohen, and Lapata 2018). It quantifies degree to which summarization models retain key information from the document. The more questions a system can answer, the better it is at summarizing the document. Following a similar paradigm, we devised a set of questions (normally 2-3 question per summary) based on the gold summary on the assumption that those summaries did, in fact, highlight the most important content in the document content. Appendix B shows some example QAs. Participants were asked to answer the given questions by only reading the summaries without access to the source articles. We elicited 3 responses per HIT. Similar with Clarke and Lapata (2010), we let the evaluation score be marked with 1 if the answer was correct, 0.5 if partially correct, and 0 otherwise.

Criteria ranking To assess the quality of the summaries, we gave the participants the full article and asked them to select the best and worst summaries from the original summary in the dataset and those produced by each model according to four specific criteria: *Informativeness* (how much useful information does the summary provide?), *Novelty* (how much new information does each summary sentence provide?), *Relevance* (how well the summary is relevant to input document?), and *Fluency* (how well the summary sentences are grammatically correct or easy to read?). We randomly select 20 instances from CNN/DM dataset to conduct the criteria ranking. The scores were computed as the percentage of times a summary was chosen as the best minus the times it was selected as the worst. The scores range from -1 (worst) to 1 (best).

Models	QA	Criteria			
		Infor.	Nov.	Rel.	Flu.
PG+Cov.	26.0*	-0.28 *	-0.43 *	-0.05 *	-0.39*
Bottom-Up	31.3*	-0.07*	0.02 *	-0.08 *	-0.02*
Inconsistency	29.8*	-0.10*	-0.12*	-0.15*	-0.14*
ESCA-BERT	39.2	0.15	0.14	0.15	0.12
Gold	♭	0.30	0.40	0.13	0.48
Bottom-Up	♭	-0.23	-0.07	-0.15	♭
ESCA-BERT	♭	**0.10**	0.03	0.05	♭
ESCA($\epsilon_n = 0.3$)	♭	0.05	**0.10**	0.02	♭
ESCA($\epsilon_r = 0.5$)	♭	0.07	-0.02	**0.07**	♭

Table 5: QA and criteria-based human evaluation. $*$ indicates statistically significant improvements over the baselines with ESCA-BERT (from a paired t-test at $p < 0.05$). Gold summaries were not included in QA evaluation. ♭ means it does not need to be evaluated by the specific use.

Based on the QA evaluation in Table 5, the summaries produced by ESCA-BERT spanned significantly more salient content. In the first block of the criteria ranking, the gold summary sets the upper bound, except for relevance. Unsurprisingly, since the gold summaries of CNN/DailyMail are mostly from the top sentences in the articles, their relevance cannot be guaranteed. We also found that ESCA-BERT produced the most popular summaries comparing with the other baseline in terms of the four criteria metrics. In the second block, ESCA with novelty and relevance controls were evaluated together with the Bottom-up and original ESCA. The difference in rankings varied slightly but, overall, the results clearly prove that ESCA with controllable novelty or relevance gained the highest rank at corresponding criteria (ϵ_n or ϵ_r, bold the highest value in the second block of Table 5).

Conclusion

This paper presents a novel hybrid framework for document summarization. The proposed ESCA is hybrid model equipped with a pair-wise ranking extractor that seamlessly connects with an abstractor armed with a sentence-level attention pointer. The flow of the framework is designed to explicitly explain why sentences are marked for extraction and to allow the operator to control exactly which sentences are ultimately extracted according to novelty and relevance scores. The subsequent abstractive generation process attends to these metrics when inferring the final summaries to produce the most desirable result. Both empirical and subjective experiments show that our model makes a statistically significant improvement over stat-of-the-art baselines.

Acknowledgments

This work was sponsored by the National Key Research and Development Program of China (2016YFB1000902), the National Natural Science Foundation of China (U19B2020, 61751201), and sponsored by CCF-Tencent Open Fund.

References

Celikyilmaz, A.; Bosselut, A.; He, X.; and Choi, Y. 2018. Deep Communicating Agents for Abstractive Summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1662–1675. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1150.

Çelikyilmaz, A.; Bosselut, A.; He, X.; and Choi, Y. 2018. Deep Communicating Agents for Abstractive Summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Volume 1 (Long Papers)*, 1662–1675.

Chen, Y.-C.; and Bansal, M. 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)*, 675–686.

Cho, J.; Seo, M.; and Hajishirzi, H. 2019. Mixture Content Selection for Diverse Sequence Generation. 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)*, 3112–3122.

Clarke, J.; and Lapata, M. 2010. Discourse Constraints for Document Compression. *Computational Linguistics* 36(3): 411–441. doi:10.1162/coli_a_00004. URL https://www.aclweb.org/anthology/J10-3005.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Durrett, G.; Berg-Kirkpatrick, T.; and Klein, D. 2016. Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1998–2008.

Gehrmann, S.; Deng, Y.; and Rush, A. 2018. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4098–4109.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, 1693–1701.

Hsu, W.-T.; Lin, C.-K.; Lee, M.-Y.; Min, K.; Tang, J.; and Sun, M. 2018. A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 132–141.

Khandelwal, U.; Levy, O.; Jurafsky, D.; Zettlemoyer, L.; and Lewis, M. 2019. Generalization through Memorization: Nearest Neighbor Language Models. *arXiv preprint arXiv:1911.00172* .

Li, W.; Xiao, X.; Lyu, Y.; and Wang, Y. 2018. Improving neural abstractive document summarization with explicit information selection modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1787–1796.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* .

Liu, Y.; and Lapata, M. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* .

Mann, W. C.; and Thompson, S. A. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text – Interdisciplinary Journal for the Study of Discourse* 8(3): 243–281.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S. J.; and McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. *arXiv preprint arXiv:2005.00661* .

Mihalcea, R.; and Tarau, P. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404–411.

Nallapati, R.; Zhai, F.; and Zhou, B. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Nallapati, R.; Zhou, B.; dos Santos, C. N.; Gülçehre, Ç.; and Xiang, B. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL*, 280–290.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Ranking Sentences for Extractive Summarization with Reinforcement Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1747–1759.

Paulus, R.; Xiong, C.; and Socher, R. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* .

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* .

Peyrard, M. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1059–1073.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Sandhaus, E. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia* 6(12): e26752.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 1: Long Papers*, 1073–1083.

Sharma, E.; Huang, L.; Hu, Z.; and Wang, L. 2019. An Entity-Driven Framework for Abstractive Summarization. *arXiv preprint arXiv:1909.02059* .

Shen, X.; Suzuki, J.; Inui, K.; Su, H.; Klakow, D.; and Sekine, S. 2019. Select and Attend: Towards Controllable Content Selection in Text Generation. 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)*, 579–590.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tan, J.; Wan, X.; and Xiao, J. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1171–1181.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, 2692–2700.

Wang, W.; Gao, Y.; Huang, H.-Y.; and Zhou, Y. 2019. Concept Pointer Network for Abstractive Summarization. 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)*, 3067–3076.

Xu, H.; Liu, B.; Shu, L.; and Yu, P. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. 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)*, 2324–2335. Association for Computational Linguistics. doi:10.18653/v1/N19-1242. URL https://www.aclweb.org/anthology/N19-1242.

You, Y.; Jia, W.; Liu, T.; and Yang, W. 2019. Improving Abstractive Document Summarization with Salient Information Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2132–2141.

Zhang, X.; Wei, F.; and Zhou, M. 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*, 5059–5069. Florence, Italy: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1499.

Zheng, H.; and Lapata, M. 2019. Sentence Centrality Revisited for Unsupervised Summarization. *arXiv preprint arXiv:1906.03508* .

Zhou, Q.; Yang, N.; Wei, F.; Huang, S.; Zhou, M.; and Zhao, T. 2018. Neural Document Summarization by Jointly Learning to Score and Select Sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 654–663.

Zhou, Q.; Yang, N.; Wei, F.; and Zhou, M. 2017. Selective Encoding for Abstractive Sentence Summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, 1095–1104.