Jointly Learning Salience and Redundancy by Adaptive Sentence Reranking for Extractive Summarization

Ximing Zhang, Ruifang Liu[*]
Beijing University of Posts and Telecommunications, Beijing 100876, China
ximingzhang@bupt.edu.cn, lrf@bupt.edu.cn

Abstract

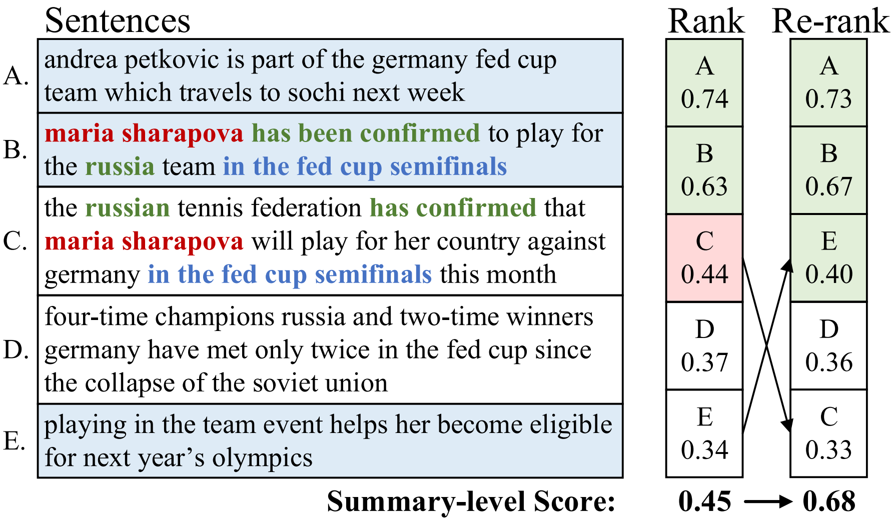
Extractive text summarization seeks to extract indicative sentences from a source document and assemble them to form a summary. Selecting salient but not redundant sentences has always been the main challenge. Unlike the previous two-stage strategies, this paper presents a unified end-to-end model, learning to rerank the sentences by modeling salience and redundancy simultaneously. Through this ranking mechanism, our method can improve the quality of the overall candidate summary by giving higher scores to sentences that can bring more novel information. We first design a summary-level measure to evaluate the cumulating gain of each candidate summaries. Then we propose an adaptive training objective to rerank the sentences aiming at obtaining a summary with a high summary-level score. The experimental results and evaluation show that our method outperforms the strong baselines on three datasets and further boosts the quality of candidate summaries, which intensely indicate the effectiveness of the proposed framework.

1 Introduction

Extractive summarization aims to create a summary by identifying and concatenating the most important sentences in a document (Nallapati et al., 2016; Zhou et al., 2018; Dong et al., 2018; Liu and Lapata, 2019; Zhang et al., 2019). To choose an appropriate sentence from the document, we need to consider two aspects: *salience*, which represents how much information the sentence carries; and *redundancy*, which represents how much information in the sentence is already included in the previously selected sentences(Peyrard, 2019). The former focuses on *sentence-level* importance, while the latter considers the relationship between sentences at the *summary-level*.

The main challenge of extractive summarization is how to combine salience and redundancy simultaneously. Most previous methods (Cheng and Lapata, 2016; Kedzie et al., 2018) only consider salience at the sentence-level, where they usually model sentence-selecting as a sequence labeling task. Several approaches for modeling redundancy between selected sentences are generally classified into two types: heuristics-based and model-based approaches(Ren et al., 2016). The former such as Trigram Blocking (Liu and Lapata, 2019) is not adaptive since they usually apply the same rule to all the documents, which results in limited effects on a few specific datasets. The latter depends heavily on feature engineering or a neural post-processing module to model redundancy. (Bi et al., 2020) extracts Ngram-matching and semantic-matching features to indicate the redundancy of a candidate sentence. (Zhong et al., 2020) proposes a two-step pipeline that first scores salience, then learns to balance salience and redundancy as a text-matching task. The critical drawbacks are error propagation and high computation cost. Compared to these models, we aim to propose an efficient and unified end-to-end model to jointly learn salience and redundancy without extra post-processing modules.

[*] Corresponding author.



Sentences

A. andrea petkovic is part of the germany fed cup team which travels to sochi next week

B. **maria sharapova** **has been confirmed** to play for the **russia** team **in the fed cup semifinals**

C. the **russian** tennis federation **has confirmed** that **maria sharapova** will play for her country against germany **in the fed cup semifinals** this month

D. four-time champions russia and two-time winners germany have met only twice in the fed cup since the collapse of the soviet union

E. playing in the team event helps her become eligible for next year's olympics

Figure 1: A summary sample whose sentences are scored by the baseline model(labeled as Rank) and our model(labeled as Re-rank) respectively. Both models only select sentences with top-3 scores as candidate summaries. Sentences on blue background constitute the gold summary. Phrases painted in the same color indicate N-gram overlap.

In this paper, we present a unified end-to-end ranking-based method for extractive summarization. Unlike previous methods, we train a single unified model that is both salience-aware and redundancy-aware to extract high-quality sentences at the summary level. The principle idea is that the best summary should consist of candidate sentences that can make the largest cumulating metric gain. As shown in Figure 1, only modeling salience when scoring sentences leads to give a high score to a salient but redundant sentence. Due to the redundancy between the sentences labeled as B and C in Figure 1 , the readers can not get the most overall information from the candidate summary consisting of the top 3 sentences. Our method aims to give higher scores to the sentences containing novel but important information by globally considering its contribution to the summary-level metric gain, and generates a candidate summary with a high summary-level score. Specifically, similar to the intuition in (Donmez et al., 2009; Ai et al., 2018), we first define a new summarization evaluation measure - Normalized Cumulating Gain based on the Overlap of N-gram (NCGON) between candidate sentences and gold summary, which can better evaluate the overall quality of candidate summaries by considering the distances of multiple N-gram matching scores between the candidate summary and golden summary. Then we design a novel redundancy-aware ranking loss - Adaptive Summary Ranking Loss (AdpSR-Loss) modified by NCGON. We penalize the deviation of the predicted ranking position probabilities of sentence pairs from the desired probabilities, which leads our model to rerank sentences adaptively according to the difference of NCGON between candidate summaries, and finally find the best candidate summary end-to-end.

We conduct experiments on a range of benchmark datasets. The experimental results demonstrate that the proposed ranking framework achieves improvements over previous strong methods on a large range of benchmark datasets. Comprehensive analyses are provided to further illustrate the performance of our method on different summary length, matching N-gram and so on.

Our contributions are as follows:

1. We propose a unified end-to-end summary-level extractive summarization model, jointly learning salience and redundancy of candidate sentences without an extra post-processing stage.

2. We consider extractive summarization as a sentence ranking task and present a new objective AdpSR-Loss based on the summary-level measure NCGON.

3. Without using extra models to reduce redundancy, we outperform the strong baseline methods by a large margin. Experimental results and analysis show the effectiveness of our approach.

2 Related Work

Neural networks have achieved great success in the task of text summarization. There are two main lines of research: abstractive and extractive. The abstractive paradigm (See et al., 2017; Narayan et al., 2018b) focuses on generating a summary word-by-word after encoding the full document. The extractive approach (Cheng and Lapata, 2016) directly selects sentences from the document to assemble into a summary. Recent research on extractive summarization spans a large range of approaches. These Extractive summarization models often use a dedicated sentence selection step aiming to address redundancy after sentence scoring step which deals with salience.

2.1 Salience Learning

With the development of neural networks, great progress has been made in extractive document summarization. Most of them focus on the encoder-decoder framework and use recurrent neural networks(Dong et al., 2018) or Transformer(Zhong et al., 2019b) encoders for the sentence scoring(Wang et al., 2020). These architectures are widely used and also extended with reinforcement learning(Zhou et al., 2018). More recently, summarization methods based on BERT (Devlin et al., 2018) have been shown to achieve state-of-the-art performance(Liu and Lapata, 2019; Zhong et al., 2019a; Zhang et al., 2019) for extractive summarization. The development of the above sentence representation and scoring models do help to achieve improvements on selecting salient sentences.

2.2 Redundancy Learning

There are relatively fewer methods that study sentence selection to avoid redundancy. In the non-neural approaches, Maximal Marginal Relevance(Carbonell, 1998) based methods select the content that has the maximal score and is minimally redundant with the previously constructed partial summary. Integer Linear Programming based methods (McDonald, 2007) formulate sentence selection as an optimizing problem under the summary length constraint. Trigram blocking (Liu and Lapata, 2019) filter out sentences that have trigram overlap with previously extracted sentences. In the neural approaches, (Zhou et al., 2018) propose to jointly learn to score and select sentences with a sequence generation model. (Bi et al., 2020) proposed redundancy-aware models by modeling salience and redundancy using neural sequence models. (Zhong et al., 2020) proposes a two-step pipeline that first scores salience, then learns to balance salience and redundancy as a text-matching task.

Compared to these methods, our method aims to propose an efficient one-stage method to jointly learn salience and redundancy without extra redundancy-aware models, and have a good generalization on a large range of benchmark datasets.

3 Method

In this section, we first introduce the overall architecture including the sentence scoring model and our training mechanism in Section 3.1. Then we introduce our designed reranking training objective in Section 3.2.

3.1 Overall Architecture

Sentence Scoring Model Given a single document consisting of sentences $[sent_1, sent_2, , sent_m]$, where $sent_i$ is the i-th sentence in the document, our task is to extract a certain number of sentences to represent the main information of source document. As shown in Figure 2, using BERT(Devlin et al., 2018) as a sentence encoder BERTEnc, we add token $[CLS]$ before each sentence and use the vector from the top BERT layer h_i as the representation of $sent_i$. To learn more inter-sentence information, several transformer layers TransEnc are used after BERT:

$$h_i = BERTEnc(sent_i) \tag{1}$$

$$h_i^L = TransEnc(h_i) \tag{2}$$

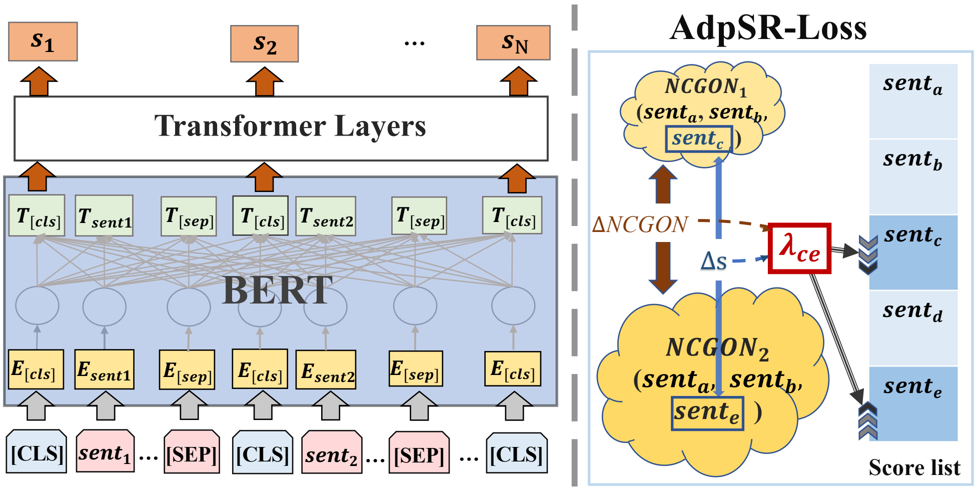


Figure 2: Overview of the proposed model. The left part is the basic sentence extraction architecture based on BERT and the right part represents the AdpSR-Loss mechanism. Δs represents the difference between the scores of $sent_c$ and $sent_e$ predicted by model in the left part.

The final output layer is a sigmoid classifier with a fully-connected layer to score each sentence:

$$\hat{y}_i = sigmoid(W_o h_i^L + b_o) \tag{3}$$

where h_i^L is the i-th representation of $sent_i$ from the transformer layers and $\hat{y}_i$ is the extraction probability of each sentence.

Training Mechanism We train the model with two different losses – binary cross entropy loss and a new adaptive ranking loss (AdpSR-Loss, Figure 2) successively. The binary-cross entropy loss aims at scoring each sentences. The novel ranking loss is designed to rerank the sentences and obtain the better sentence combination. Using this strategy, the summary-level gain of top k sentences can be efficiently improved end-to-end without introducing a second-stage model.

3.2 Extraction Summarization as Ranking

3.2.1 NCGON: A Summary-level Evaluation Measure

We refer to $D = \{sent_n | n \in (1, N)\}$ as a single document consisting of n sentences and $C = \{sent_k | sent_k \in D\}$ as a candidate summary consisting of k sentences extracted from D. A sequence of labels $\{y_n | n \in (1, N)\}$ ($y_n \in (0, 1)$) for $\{sent_n\}$ are given as ground truth by a greedy algorithm similar to (Nallapati et al., 2016) by maximizing the ROUGE score against the gold summary C^* written by human. Let $\{s_n | n \in (1, N)\}$ represent the scores of sentences $\{sent_n\}$ predicted by model. Traditional method simply ranks $\{sent_n\}$ according to $\{s_n\}$ in descending order and selects top k sentence as candidate summary C^t, and the remaining sentences are abandoned directly by model, which does not consider the overall gain for obtaining summaries.

To better evaluate the gain of sentence choosing at a summary-level, we consider the overlap between candidate and gold summary as cumulating gain (CG). As Figure 1 shows, the CG of C^t may not be the highest due to the redundancy, although the score of each $sent_i$ is top k highest individually. We use function as followed to measure the overlap of N-grams (N=2,3,4) between the extracted summary C^t and the gold summary C^*:

$$R_n(C^*, C^t) = Rouge\text{-}N(C^*, C^t) \tag{4}$$

Consequently, the cumulating gain based on the overlap of N-grams (CGON) as follows:

$$CGON = \sum_{n=2,3,4} R_n(C^*, C^t) \tag{5}$$

Swapping the position of $sent_c$ and $sent_e$ in Figure 1 makes CGON improve by minimizing the overlap between top k sentences. Therefore, using CGON as the measure, we can better find the summary-level candidate containing most overall information rather than only using cross entropy loss which tends to select sentences with higher individual gain. To fairly quantify the CGON of different candidates, we normalize it by $CGON_{max}$ as our final evaluation measure NCGON:

$$NCGON = \frac{CGON}{CGON_{max}} \tag{6}$$

where $CGON_{max}$ is the CGON of the ground truth candidate consisting of sentences whose labels are equal to 1.

3.2.2 AdpSR-Loss: Adaptive Summary Ranking Loss

Inspired by the Learning to Rank(LTR) structure (Burges et al., 2005; Donmez et al., 2009), we model extractive summarization as a pair-wise ranking problem and aim at reranking the sentence list to find the best candidate summary consisting of sentences with top k scores. Based on the method in 2.1, we get the original score list using cross entropy loss. Considering one pair of sentences $\{sent_i, sent_j\}$, where $i \in (1, k)$ and $j \in (k+1, N)$. Let $U_i > U_j$ denote that $sent_i$ is ranked higher than $sent_j$. $\{s_i, s_j\}$ is mapped to a learned probability P_{ij} as followed, which indicates the probability of ranking position between $sent_i$ and $sent_j$ predicted by model.

$$P_{ij} = P(U_i > U_j) = \frac{1}{1 + e^{-(s_i - s_j)}} \tag{7}$$

We apply the cross entropy loss, which penalizes the deviation of the model output probabilities P_{ij} from the desired probabilities $\overline{P}_{ij}$: let $\overline{P}_{ij} \in \{0, 1\}$ be the ground truth of the ranking position of $\{sent_i, sent_j\}$. Then the cost is:

$$\begin{aligned} L_{ij} &= -\overline{P}_{ij} log P_{ij} - (1 - \overline{P}_{ij}) log(1 - P_{ij}) \\ &= log(1 + e^{-(s_i - s_j)}) \end{aligned} \tag{8}$$

To better optimize the evaluation measure NCGON through L_{ij}, we modify L_{ij} by simply multiplying ΔNCGON, which represents the size of the change in NCGON given by swapping the rank positions of $sent_i$ and $sent_j$ (while leaving the rank positions of all other sentences unchanged). Specific to extractive summarization, there is no need to calculate all pairs of sentences in D, we only focus on $\{sent_i,\ sent_j | i \in (1, k), j \in (k + 1, N)\}$ due to the fact that only the overlap of N-grams between top k sentences in the ranking list and gold summary make sense. Finally, the Adaptive Summary Ranking Loss (AdpSR-Loss) can be written as:

$$L = \sum_{i,j} L_{ij} \quad = \sum_{i,j} log(1 + e^{-(s_i - s_j)}) |\Delta NCGON| \tag{9}$$

Our experiments have shown that such a ranking loss actually optimizes NCGON directly, which leads our model to extract a better candidate summary.

3.2.3 Understanding How AdpSR-Loss Works

To explain AdpSR-Loss commonly, we define the gradient of the cost L_{ij} as $|\lambda_{ij}|$, which we can easily get through derivation:

$$\lambda_{ij} = -\frac{1}{1 + e^{(s_i - s_j)}} |\Delta NCGON| \tag{10}$$

$|\lambda_{ij}|$ can be interpreted as a force: if $sent_j$ is more salient than $sent_i$, which means choosing $sent_j$ could obtain higher cumulating gain than $sent_i$, then $sent_j$ will get a push upwards of size $|\lambda_{ij}|$. By multiplying $\Delta NCGON$, the gradient is endowed with practical physical meaning: sentences that can bring more relative gain will be given greater power to improve their ranking position adaptively.

4 Experiments

4.1 Datasets

We conduct empirical studies on three benchmark single-document summarization datasets, CNN/DailyMail (Hermann et al., 2015), Xsum (Narayan et al., 2018a) and WikiHow (Koupaee and Wang, 2018) as followed.

CNN/DailyMail is a widely used summarization dataset for single-document summarization, which contains news articles and associated highlights as summaries.

XSum is a one-sentence summary dataset to answer the question "What is the article about?". We use the splits of Narayan et al. (2018a) for training, validation, and testing.

WikiHow is a new large-scale dataset using the online WikiHow knowledge base.

Table 1 shows the full statistics of three datasets. All the sentences are split with the Stanford CoreNLP toolkit (Manning et al., 2014) and pre-processed following (Liu and Lapata, 2019). We tokenize sentences into subword tokens, and truncate documents to 512 tokens.

Datasets	**Source**	**# Pairs**			**# Tokens**		**# Ext**
		Train	Valid	Test	Doc.	Sum.	
CNN/DM	News	287,084	13,367	11,489	766.1	58.2	3
XSum	News	203,028	11,273	11,332	430.2	23.3	2
WikiHow	KB	168,126	6,000	6,000	580.8	62.6	4

Table 1: Datasets overview. The data in Doc. and Sum. indicates the average length of documents and summaries in the test set respectively. # Ext denotes the number of sentences that should extract in different datasets.

4.2 Implementation Details

Our baseline BertSum (Liu and Lapata, 2019) using BERT(Devlin et al., 2018) as a sentence encoder, the vectors from the top BERT layer of token $[CLS]$ before each sentence are used as the representation of each sentence. To learn more inter-sentence information, several transformer layers are used after BERT and a sigmoid classifier is stacked to score each sentence. In order to avoid interference of other factors, we re-implement the model BERTSUM in our training environment according to the default parameters of (Liu and Lapata, 2019) using the base version of BERT[0], and compare our method on this baseline fairly. All the models are trained on 2GPUs (GTX 1080 Ti) with gradient accumulation per two steps. We use Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.999$ and adopt the learning rate schedule as (Vaswani et al., 2017) with a warming-up strategy. We train the model with cross entropy loss for 50,000 steps to obtain the original score for each sentence and AdpSR-Loss afterward with max learning rate $2e^{-5}$ for 3000 steps to obtain the final scores.

4.3 Evaluation Metric

We adopt ROUGE (Lin, 2004) for evaluation metric, which is the standard evaluation metric for summarization. We report results in terms of unigram and bigram overlap (ROUGE-1 and ROUGE-2) as a means of assessing informativeness, and the longest common subsequence (ROUGE-L) as a means of assessing fluency.

4.4 Experimental Results

Table 2 summarizes the results of CNN/DM dataset using ROUGE-F1 evaluation. The first block in the table includes the results of an extractive ORACLE system as an upper bound and a LEAD-3 baseline (which simply selects the first three sentences in a document). The second block summarizes the strong

[0] https://github.com/huggingface/pytorch-pretrained-BERT

Model	R-1	R-2	R-L
LEAD	40.43	17.62	36.67
ORACLE	52.59	31.23	48.87
BanditSum* (Dong et al., 2018)	41.50	18.70	37.60
NeuSum* (Zhou et al., 2018)	41.59	19.01	37.98
HiBert* (Zhang et al., 2019)	42.37	19.95	38.83
BertExt* (Bae et al., 2019)	42.29	19.38	38.63
BertSum(Liu and Lapata, 2019)	42.54	19.86	39.00
BertSum + Tri-Blocking	42.86(+0.32)	19.87(+0.01)	39.29(+0.29)
BertSum + Reranking (Ours)	**42.94(+0.40)**	**20.04(+0.18)**	**39.31(+0.31)**

Table 2: Results on CNN/DM test set. Results with ∗ mark are taken from the corresponding papers.

extractive summarization baselines on CNN/DM. The third block shows our proposed method results on R-1, R-2 and R-L compared to BERTSUM and BERTSUM without Tri-blocking. Compared with BERTSUM without Tri-blocking, our method achieves 0.40/0.18/0.31 improvements on R-1, R-2 and R-L. Also, we outperforms BERTSUM with Tri-blocking, which is the most commonly used and effective method to remove redundancy on CNN/DM.

Table 3 presents results on the XSum and WikiHow dataset. Our model achieves 0.49/0.13/0.75 improvements on R-1, R-2, and R-L in the Xsum dataset and 1.56/0.10/1.31 improvements on R-1, R-2, and R-L in the WikiHow dataset. Notably, using Tri-Blocking on these two datasets leads to a decrease in performance. Compared with using Tri-Blocking for redundancy removal, our method has improvements on all the three datasets, which illustrates that the reranking mechanism has better generalization ability on summary-level sentence extracting systems.

In addition, we find the scores improve especially on XSum and WikiHow by a large margin. And the baseline model tends to choose longer sentences than our model. Through calculation, we can get the average sentence length (19.5/11.65/15.6) of the three datasets (CNNDM/XSum/ WikiHow), which indicates our summary-level model is more powerful than the sentence-level framework, especially when the gold summaries consist of shorter sentences.

Model	R-1	R-2	R-L
XSum (Num = 2)			
LEAD	16.30	1.60	11.95
ORACLE	29.79	8.81	22.66
BERTSUM(Liu and Lapata, 2019)	22.83	4.38	16.96
BERTSUM + Tri-Blocking	22.72(-0.11)	4.18(-0.20)	17.21(+0.25)
BERTSUM + Reranking (Ours)	**23.32(+0.49)**	**4.51(+0.13)**	**17.71(+0.75)**
WikiHow (Num = 4)			
LEAD	24.97	5.83	23.24
ORACLE	35.59	12.98	32.68
BERTSUM(Liu and Lapata, 2019)	30.08	8.39	28.00
BERTSUM + Tri-Blocking	30.00(-0.08)	8.25(-0.14)	27.95(-0.05)
BERTSUM + Reranking (Ours)	**31.64(+1.56)**	**8.49(+0.10)**	**29.31(+1.31)**

Table 3: Results on test sets of WikiHow and XSum. Num indicates how many sentences are extracted as a summary.

5 Qualitative Analysis

To further analyze the main results in Section 4.4, we carry out detailed evaluation and analysis. We first give an ablation study of our method in Section 5.1 , and then analyze the performance of our method on different summary length in Section 5.2. Considering that the main metric ROUGE is based on the N-gram overlap, we conduct a fine-grained analysis in Section 5.3. Finally, we implement a human evaluation and case study in Section 5.4 and 5.5.

5.1 Ablation Study

To study the effectiveness of each component of the AdpSR-Loss, we conduct several ablation experiments on the CNN/DM dataset. "w/o R-N" denotes that we remove the Rouge-N gain and only use the sum of remaining gain to weight the AdpSR-loss. As the results shown in Table 4, we could find R-2 is an essential element in NCGON to increase the outcome of our strategy, comparing to R-3 and R-4 that we guessed in advance. However, increasing the weight of R-2 in NCGON can not bring better results.

Model	R-1	R-2	R-L
OUR METHOD	**42.94**	**20.04**	**39.31**
W/O R-2	42.69	19.96	39.13
W/O R-3	42.80	20.00	39.23
W/O R-4	42.83	20.00	39.21
W/O (R-3+R-4)	42.70	19.98	39.21
W/O (R-2+R-4)	42.65	19.96	39.09

Table 4: Results of removing different components of our ranking loss on the CNN/DM dataset.

5.2 Effect of Summary Length

To analyze the performance of our method on different summary length, we divide the test set of CNN/DM into 5 intervals based on the length of gold summaries (X-axis in Figure 3). We evaluate the performance of our method and the baseline BERTSUM in various parts, and the improvements of the sum of scores on R-1, R-2, R-L are drawn as bars (left y-axis ΔR). As shown in Figure 3, the ROUGE increases more significantly on documents with short summaries, which means our model can efficiently extract short but salient sentences instead of long sentences prone to redundancy. This further proves that the joint consideration of salience and redundancy during model training is effective on obtaining extractive summaries, especially on summaries consisting of short sentences.

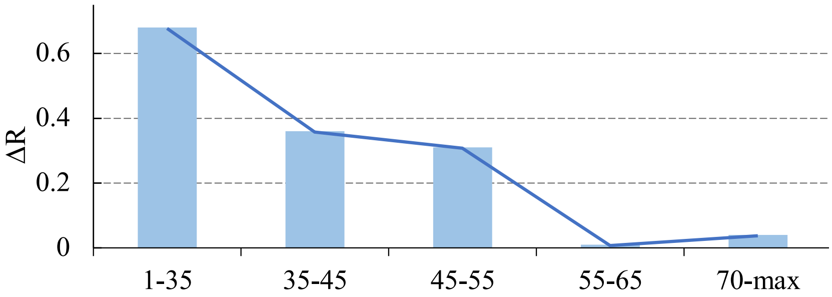


Figure 3: Datasets splitting experiment. The X-axis indicates the length of gold summaries, and the Y-axis represents the ROUGE improvement of OUR MODEL over BERTSUM on this subset.

5.3 Analysis of N-gram Frequency

To further analyze the difference between system summary and the oracle, we count the n-gram frequency in the source document of the matching n-gram and unmatching n-gram between system summary and oracle in Table 5 on CNN/DM. Here F-match means the n-gram frequency of the matching n-grams and F-unmatch means the n-gram frequency of the unmatching n-grams.

Datasets	F-match	F-unmatch
BERTSUM	5.53	1.68
BERTSUM + TRI-BLOCKING	5.52	1.68
BERTSUM + RERANKING	5.50	1.69

Table 5: Results of the n-gram frequency of BertSum on the CNN/DM and WikiHow test set.

Compared with matching n-grams, the average n-gram frequency of unmatching n-grams is much lower, and the frequency is almost the same in all the comparing models (the frequency of unmatching n-grams on the three models are similar to equal), which shows that finding sentences containing perl but important n-grams with lower n-gram frequency is an important aspect of improving the summary quality.

5.4 Human Evaluation

As we know, although the ROUGE metric has long been regarded as a classical evaluation metric in summarization, it does not always reflect the quality of salience and redundancy. Hence, we conduct human evaluation to further analyze. We follow the human evaluation method proposed in (Zhou et al., 2018), which is widely used in the extractive summarization task. We randomly sample 200 documents and ask five volunteers to evaluate the summaries of the two model outputs. They rank the output summaries as 1 (best) or 2 (worst) regarding informativeness, redundancy and overall quality, and they evaluate the summaries by the fair and anonymous ranking method. Table 6 shows the human evaluation results. Our method performs better than the baseline model BERTSUM, especially in Rdnd, which demonstrates our model is more redundancy-aware.

Model	Info	Rdnd	Overall
BERTSUM	1.56	1.64	1.59
BERTSUM + RERANKING	**1.44**	**1.36**	**1.41**

Table 6: Average ranks of BERTSUM and our method on CNN/DM in terms of informativeness (Info), redundancy (Rdnd) and overall quality by human (lower is better).

5.5 Case Study

We investigate two examples of extracted output in Figure 4. As the first case illustrates, comparing to BERTSUM, our reranking method effectively avoids the n-gram overlap between the chosen sentences, and chooses more correct sentences, which means our method could find the sentence combination with less redundancy and more salient information. Also, we also find that during evaluating BERTSUM and our method, both of these two systems tend to choose sentences with similar positions. As the cases show, BERTSUM chooses the sentences at position 0, 1 in the first case and position 5, 6 in the second case. Besides, BERTSUM also tends to choose sentences that are closer to the beginning of the article. Comparing to BERTSUM, our reranking method alleviates this problem to a certain extent, but the trend still exists, which may also have a correlation with the characteristics of the CNN/DM dataset itself.

Article #1 (label: 0, 2, 15)	**Candidate #1**
john carver says his newcastle players have a point to prove to themselves at liverpool on monday night . the magpies are in danger of being sucked into what had previously seemed an unlikely relegation battle given their alarming run of form . amid accusations of the team ` playing with their flip-flops on ` , united have lost four on the spin - scoring just once - and carver admits he does not know where their next point is coming from . it is unlikely to arrive at anfield , a venue at which they last won in 1994 . they were beaten 1-0 by north-east rivals sunderland seven days ago -- a performance carver labelled 'em barrassing ' -- and there was a showdown meeting at the club 's training ground this week in which the head coach let his feelings be known . carver , though , is hoping that personal pride will kick in when they travel to Merseyside … i have to accept what people say . as long as it is done the right way and in a constructive manner , i can accept that , ' he said . newcastle were beaten in last weekend 's tyne/wear derby by jermain defoe 's stunning volley. carver has seen his newcastle team lose four in a row and fall into relegation danger …	**BERTSUM (candidate: 0, 1, 7)** • john carver says his newcastle players have a point to prove to themselves at liverpool on monday night . • john carver says his newcastle players have a point to prove when they face liverpool on monday. • the magpies are in danger of being sucked into what had previously seemed an unlikely relegation battle given their alarming run of form . **OURS (candidate: 0, 1, 15)** • john carver says his newcastle players have a point to prove to themselves at liverpool on monday night . • the magpies are in danger of being sucked into what had previously seemed an unlikely relegation battle given their alarming run of form • newcastle were beaten in last weekend 's tyne/wear derby by jermain defoe 's stunning volley.
Article #2 (label: 3, 4, 9)	**Candidate #2**
two hours before the miami open semifinal , novak djokovic practiced his returns in an empty stadium …` but i managed to get a lot of serves back . that was one of the keys in the match , making him play and getting into the rally and making him work extra . novak djokovic beat john isner in straight sets to reach the finalof the miami open on friday night . the no 1-seeded djokovic closed to within one win of his fifth key biscayne title and will face andy murray . the no 1-seeded djokovic closed to within one win of his fifth key biscayne title . his opponent sunday will be two-time champion andy murray... djokovic 's biggest hole while serving was a love-30 deficit late in the first set . he responded with consecutive aces and escaped . djokovic is aiming to win his fifth title in miami and will take on scotsman murray in sunday 's final . djokovic 's first break gave him a 2-1 edge in the second set , and that margin grew to 5-1 . he finished with just eight unforced to 31 by isner , who lost 70 percent of his second-serve points …	**BERTSUM (candidate: 3, 5, 6)** • the no 1-seeded djokovic closed to within one win of his fifth key biscayne title and will face andy murray. • novak djokovic beat john isner in straight sets to reach the finalof the miami open on friday night. • his opponent sunday will be two-time champion andy murray , who defeated tomas berdych 6-4 , 6-4 . **OURS (candidate: 3, 4, 9)** • novak djokovic beat john isner in straight sets to reasch the finalof the miami open on friday night . • the no 1-seeded djokovic closed to within one win of his fifth key biscayne title and will face andy murray. • djokovic is aiming to win his fifth title in miami and will take on scotsman murray in sunday 's final.

Figure 4: Example output articles, candidate summaries from BERTSUM and our method from CNN/DM dataset. Sentences painted in green color are the golden sentences. Phrases painted with the grey highlight indicate N-gram overlaps.

6 Conclusions

In this paper, we propose a novel end-to-end summary-level method jointly learning salience and redundancy in the extractive summarization task. Experiments on three benchmark datasets confirm the effectiveness of our one-stage mechanism based on the adaptive ranking objective. Moreover, we find that our method delivers more benefits to short-sentence summaries. We believe the power of this ranking-based summarization framework has not been fully exploited especially on the design of cumulating gain and fine-grained learning, and we hope to provide new guidance for future summarization work.

Acknowledgements

We would like to thank the CCL reviewers for their valuable comments and Keqing He, Pengyao Yi and Rui Pan for their generous help and discussion.

References

Qingyao Ai, Keping Bi, Jiafeng Guo, and W Bruce Croft. 2018. Learning a deep listwise context model for ranking refinement. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 135–144.

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sang-goo Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 10–20, Hong Kong, China, November. Association for Computational Linguistics.

Keping Bi, R. Jha, W. Croft, and A. Çelikyilmaz. 2020. Aredsum: Adaptive redundancy-aware iterative sentence ranking for extractive document summarization. *ArXiv*, abs/2004.06176.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96.

J Carbonell. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998*, pages 335–336.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany, August. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748.

Pinar Donmez, Krysta M Svore, and Christopher JC Burges. 2009. On the local optimality of lambdarank. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 460–467.

K. Hermann, Tomás Kociský, E. Grefenstette, Lasse Espeholt, W. Kay, Mustafa Suleyman, and P. Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

C. Kedzie, K. McKeown, and Hal Daumé. 2018. Content selection in deep learning models of summarization. In *EMNLP*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

M. Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *ArXiv*, abs/1810.09305.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Yang Liu and M. Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP/IJCNLP*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, pages 557–564. Springer.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.

S. Narayan, Shay B. Cohen, and M. Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

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

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

Pengjie Ren, Furu Wei, Zhumin Chen, Jun Ma, and Ming Zhou. 2016. A redundancy-aware sentence regression framework for extractive summarization. In *COLING*, pages 33–43.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, A. Gomez, L. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuan-Jing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219.

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

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2019a. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058.

Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2019b. A closer look at data bias in neural extractive summarization models.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and X. Huang. 2020. Extractive summarization as text matching. *ArXiv*, abs/2004.08795.

Qingyu Zhou, Nan Yang, F. Wei, Shaohan Huang, M. Zhou, and T. Zhao. 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)*.