Improving Unsupervised Extractive Summarization with Facet-Aware Modeling

Xinnian Liang[1]*, **Shuangzhi Wu**[2], **Mu Li**[2] and **Zhoujun Li**[1]†
[1]State Key Lab of Software Development Environment, Beihang University, Beijing, China
[2]Tencent Cloud Xiaowei, Beijing, China
{xnliang,lizj}@buaa.edu.cn; {frostwu,ethanlli}@tencent.com;

Abstract

Unsupervised extractive summarization aims to extract salient sentences from documents without labeled corpus. Existing methods are mostly graph-based by computing sentence centrality. These methods usually tend to select sentences within the same facet, however, which often leads to the facet bias problem especially when the document has multiple facets (i.e. long-document and multi-documents). To address this problem, we proposed a novel facet-aware centrality-based ranking model. We let the model pay more attention to different facets by introducing a sentence-document weight. The weight is added to the sentence centrality score. We evaluate our method on a wide range of summarization tasks that include 8 representative benchmark datasets. Experimental results show that our method consistently outperforms strong baselines especially in long- and multi-document scenarios and even performs comparably to some supervised models. Extensive analyses confirm that the performance gains come from alleviating the facet bias problem.

1 Introduction

Document summarization is the task of transforming a long document into a shorter version while retaining its most important content (Nenkova and McKeown, 2011).Existing extractive or abstractive methods are mostly in supervised fashion which rely on large amounts of labeled corpora (Cheng and Lapata, 2016; Nallapati et al., 2017; Gehrmann et al., 2018; Liu and Lapata, 2019a,b; Zhang et al., 2019; Wang et al., 2020). However, this is not available for different summarization styles, domains, and languages. Fortunately, recent work has shown successful practices on unsupervised extractive summarization (Radev et al., 2000; Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Schluter and Søgaard, 2015; Tixier et al., 2017; Zheng and Lapata, 2019; Xu et al., 2020; Dong et al., 2020). Compare with supervised ones, unsupervised methods 1). remove the dependency on large-scale annotated document-summary pairs; 2). are more general for various scenarios.

Graph-based models are commonly used in unsupervised extractive methods (Radev et al., 2000; Mihalcea and Tarau, 2004; Erkan and Radev, 2004). For example, Zheng and Lapata (2019) proposed a directed centrality-based method named PacSum

Document

Facet 1: Lampard was fired.
1. As Chelsea's winter had turned bleak, as whispers that Lampard, its inexperienced coach, might be drifting toward the edge grew louder …
2. Not for the manager — that Lampard was fired so soon after he was given such public backing illustrates, quite neatly, how little power fans have — but for the public itself.
3. Chelsea might, in truth, have fired Lampard earlier. His colleagues, certainly, have been fearing it for weeks.
4. That was Sunday afternoon. He was fired on Monday morning …

Facet 2: Fans support Lampard.
5. "In Frank We Trust," it read, white letters on a blue field … And underneath, three simple words: "Then. Now. Forever." …
6. But none — not even Mourinho — have retained the support of the fans quite so unanimously as Lampard.
7. Lampard's association with Chelsea runs long and deep enough that he has deep-seated, well-established connections with Chelsea's fans.

Facet 3: Many managers were fired of Abramovich's ruthless impatience.
8. Lampard was not the first manager at Roman Abramovich's Chelsea to come under what seemed, on the surface, to be an undue, premature sort of pressure. …
9. Some of those who have gone before Lampard have done so with sympathy, perceived as victims of Abramovich's ruthless impatience.

Baseline

2. Not for the manager — that Lampard was fired so soon after he was given such public backing illustrates, quite neatly, how little power fans have — but for the public itself.
3. Chelsea might, in truth, have fired Lampard earlier. His colleagues, certainly, have been fearing it for weeks.
4. That was Sunday afternoon. He was fired on Monday morning …

Gold Reference

3. Chelsea might, in truth, have fired Lampard earlier. His colleagues, certainly, have been fearing it for weeks.
6. But none — not even Mourinho — have retained the support of the fans quite so unanimously as Lampard.
9. Some of those who have gone before Lampard have done so with sympathy, perceived as victims of Abramovich's ruthless impatience.

Figure 1: Examples from New York Times. We selected part of key sentences from the source document to show in this table. "..." refers to the omissions of context sentences due to space limitation.

*Contribution during internship at Tencent Inc.
†Corresponding Author

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1685–1697

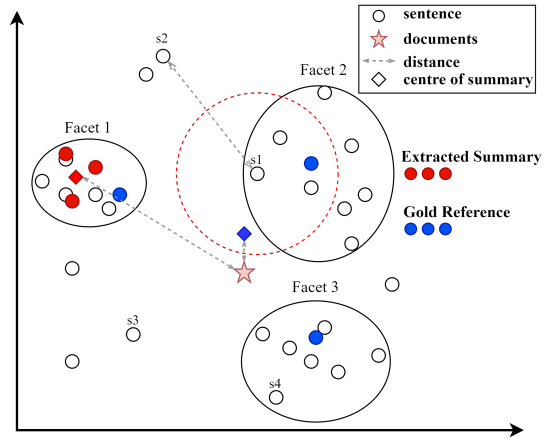

Figure 2: Visualization of facet bias. Nodes refer to sentence representations and star is the document representation. Black solid circles mean facets. Red dashed circle means threshold in Section 3.1. The dashed bi-direction arrows denote the sentence similarities.

by assuming that the contribution of any two nodes to their respective centrality is influenced by their relative position in a document. Dong et al. (2020) further improved PacSum by incorporating hierarchical and positional information into the directed centrality method. The core idea of centrality-based models is that the more similar a sentence is to other sentences, the more important it is (Radev et al., 2000). This usually works well for documents with a single facet (i.e. topic, aspect). However, there is always more than one facet, especially in long-document or multi-documents. Figure 1 shows an example of a long-document with 3 facets. We highlight the key phrases of each facet in different colors. Current centrality-based models often select sentences from one facet which is supported by more similar sentences. For example, the baseline model selects 3 sentences from facet 1. We call this the facet bias problem.

Figure 2 shows an intuitive explanation of the facet bias problem. The nodes are sentence representations, the star is the document representation and rhombuses are the centers of selected summary sentences. The sentences that support the same facet are masked in the same circle. Centrality-based models tend to select sentences from facet 1 (red nodes). Because these sentences are more similar to each other which leads to a higher centrality score. However, the true summary should consist of important sentences from different facets (blue nodes). To address the facet bias problem, in this paper, we proposed a facet-aware centrality-based model, which is called Facet-Aware Rank (FAR). First, we introduce a modified graph-based ranking method to filter irrelevant sentences. Then we encode the whole document into vector space which is used to capture all facets in the document. For each candidate summary, we calculate a similarity score between the summary sentences and the document. This sentence-document similarity aims at measuring the relevance between summary and document. Whereas the sentence centrality measures the sentence-level importance. In the ranking phase, we combine the sentence-document similarity and the sentence centrality to guarantee the selected sentences are important and cover all facets. As shown in Figure 2, by incorporating the sentence-document similarity, we are more likely to select the blue ones, that is closer to the star, instead of the red ones. We evaluate our method on 8 representative datasets. The results show that our model can surpass strong unsupervised baselines on most datasets and is comparable to supervised models on some datasets. Extensive analyses confirm that the performance gains indeed come from alleviating the facet bias problem. Besides, we surprisedly find that our method can tackle redundancy in summary to some extent.

2 Background: Graph-based Ranking

Given a document D, it contains a set of sentences $\{s_1, .., s_i, .., s_j, .., s_n\}$. Graph-based algorithms treats D as a graph $\mathcal{G} = (V, E)$. $V = \{v_1, v_2, \ldots, v_n\}$ is the vertex set where v_i is the representation of sentence s_i. E is the edge set, which is an $n \times n$ matrix. Each $= \{e_{i,j}\} \in E$ denotes the weight between vertex v_i and v_j.

The key idea of graph-based ranking is to calculate the centrality score of each sentence (or vertex). Traditionally, this score is measured by degree or ranking algorithms (Mihalcea and Tarau, 2004; Erkan and Radev, 2004) based on PageRank (Brin and Page, 1998). Then the sentences with the top score are extracted as a summary. The undirected graph algorithm compute the sentence centrality score as follows:

$$\texttt{Centrality}(s_i) = \sum_{j=1}^{N} e_{ij} \qquad (1)$$

This is based on the assumption that the contribution of the sentence's importance in the document is not affected by the order of the sentence. In contrast, directed graph-based ranking algorithm takes the positional information into consideration,

which is based on the assumption that the previous content of current sentence and the later contexts have different impact on current sentence's centrality score (Mann and Thompson, 1988). Then equation 1 is reformulated as

$$\texttt{DC}(s_i) = \lambda_1 \sum_{i>j} e_{ij} + \lambda_2 \sum_{i<j} e_{ij} \quad (2)$$

Where $\lambda_1 + \lambda_2 = 1$. Hyper-parameters λ_1 and λ_2 were used to adjust the influence of previous and last content. Our method is built based on the directed graph-based ranking algorithm.

3 Facet-Aware Centrality-based Model

3.1 Modified Directed Graph-based Ranking

We propose a variation of directed graph-based ranking in this section. We modify Equation 2 in terms of filtering negligible sentences. We take s_1 in Figure 2 as an example to give an intuitive explanation. There usually exist many unrelated sentences especially in long documents for s_1 i.e. s_2, s_3, s_4. As shown in equation 2, all these sentences have a contribution in computing s_1's centrality score. We regard sentences like them as noise of s_1 and propose a modified directed graph-based ranking to filter them. To this end, we simply introduce a threshold ϵ to Equation 2. For s_1, ϵ can be seen as a diameter, s_1 is the centre. The centrality score of s_1 only consider nodes in red dashed circle. We further rewrite 2 as :

$$\begin{aligned}\texttt{DC}(s_i) =& \lambda_1 \sum_{i>j} \texttt{Max}((e_{ij} - \epsilon), 0) \\ &+ \lambda_2 \sum_{i<j} \texttt{Max}((e_{ij} - \epsilon), 0)\end{aligned} \quad (3)$$

where $\epsilon = \beta \cdot (\max(e_{ij}) - \min(e_{ij}))$. β is a Hyper-parameter to control the scale of diameter. As shown in Equation 3, if the similarity between s_i and s_j is lower than ϵ, s_j is neglected. We find this modification is very effective but the model is very sensitive to the selection of β, so we carefully tune β on the development set. We finally rank and select sentences with Equation 4.

$$\texttt{summary} = \texttt{topK}(\{\texttt{DC}(s_i)\}_{i=1,..,n}) \quad (4)$$

Where top-ranked k sentences will be extracted as summary and k is pre-defined with the average length of summary in training data.

3.2 Facet-Aware Centrality Scoring

In this section, we introduce how to implement Equation 3 and how we incorporate facet into centrality-based ranking in detail. We propose a simple method to model the facets in a document by a special representation based on the whole document.

Specifically, based on Equation 4, we add a sentence-document similarity, which computes the similarity between sentences in candidate summary and document to measure the relevance between summary $\mathcal{C}$ and document d. Candidate summary is pre-selected sentences from top-ranked K sentences with score $\texttt{DC}(s_i)$ to reduce search range. We combine sentence-document similarity with sentence centrality and obtain the best candidate summary by 5.

$$\texttt{summary} = \texttt{arg}\max_{\mathcal{C}}(\texttt{sim}(d, \hat{v}) \cdot \sum_{s_i \in \mathcal{C}} \texttt{DC}(s_i)^{\alpha}) \quad (5)$$

where α is a hyper-parameter to control the influence of directed centrality. $\texttt{sim}(d, \hat{v})$ refers to the sentence-document similarity, where d is the document representation and $\hat{v}$ is the candidate summary representation. $\hat{v}$ is obtained by $\frac{\sum_{i \in \mathcal{C}}(v_i)}{|\mathcal{C}|}$ which is the mean representation of summary sentences. We select the cosine similarity for $sim(\cdot)$.

The combination of sentence-document similarity and sentence centrality can not only tackle the facet problem but also reduce the redundancy to some extent. As shown in Figure 2, the centrality score of red nodes is extremely high due to they are similar to each other. Previous centrality-based models tend to select them as the summary. We incorporate document representation and sentence-document similarity to weight centrality score. This force model chooses the blue nodes, whose center is closer to the star, instead of red nodes. The introduction of sentence-document similarity makes it extremely unlikely that nodes of high cohesion will be selected. Thus, the redundancy is also reduced.

A candidate summary $\mathcal{C}$ is the subset of top-ranked K sentences after ranking with $\texttt{DC}(s_i)$, which satisfy the following two conditions: 1) the length of sentences in candidate summary is pre-defined L, which is related to the summary length of dataset training data; 2) the total length of top-ranked K sentences is $t \times L$, where t is empirically set as 3. For the sentence representations v_i, we employ BERT as encoder which maps each word into

a hidden state. Specifically, the sentence representations v_i is obtained by $\texttt{sigmoid}(h_i)$, where h_i is the hidden state of "[CLS]". Each e_{ij} in E is calculated by the dot product of the two sentences $v_i^\top v_j$. For document representation, we first collect all the sentence representations $\{v_1, v_2, \ldots, v_n\}$. To compress all the valuable information in the document, we apply a maxpooling function to sentence representations. The document representation d is computed as

$$d = \texttt{Maxpooling}(\{v_1, v_2, \ldots, v_n\}) \quad (6)$$

3.3 Improved Sentence Representation

The sentence representations plays a crucial role in our ranking model. The previous study shows that improving the quality of sentence representations helps improve the ranking performance (Zheng and Lapata, 2019; Dong et al., 2020). We post-train BERT on a sentence-level task constructed based on the corpus of a specific task. The idea is that its representation is affected not only by the words in it, but also the sentences around it. For a sentence in a document, we take its previous sentence and its following sentence to be positive examples and random sample sentences from documents as negative examples. The objective function follows that used in (Reimers and Gurevych, 2019). Specifically, for sentence s_i, a positive sentence s_j, and a negative sentence s_k, the BERT is trained to minimize the following equation:

$$\max(\parallel v_i - v_j \parallel - \parallel v_i - v_k \parallel + \mu, 0) \quad (7)$$

where v is the sentence representation, and μ is margin which ensures that v_j is at least μ closer to s_i than s_k. The hidden state vector of "[CLS]" is used as sentence representations and we set μ to 1 following (Reimers and Gurevych, 2019) in post-training phase.

4 Experiments

4.1 Datasets

We introduce the datasets used in our experiments in this section.

CNN/DM dataset contains 93k articles from CNN, and 220k articles from Daily Mail newspapers (Hermann et al., 2015). We use the non-anonymous version. Following (Zheng and Lapata, 2019), documents whose length of summaries are shorter than 30 tokens are filtered out.

NYT dataset contains articles published by the New York Times between January 1, 1987 and June 19, 2007 (Li et al., 2016). The summaries are written by library scientists. Different from CNNDM, salient sentences distribute evenly in an article (Durrett et al., 2016). We filter out documents whose length of summaries are shorter than 50 tokens (Zheng and Lapata, 2019).

MultiNews dataset consists of news articles and human-written summaries. The dataset is the first large-scale Multi-Documents Summarization (MDS) news dataset and comes from a diverse set of news sources (over 1500 sites) (Fabbri et al., 2019).

arXiv&PubMed datasets are two long document datasets of scientific publications from arXiv.org (113k) and PubMed (215k) (Cohan et al., 2018). The task is to generate the abstract from the paper body.

WikiSum dataset is a multi-documents summarization dataset from Wikipedia (Liu et al., 2018). We use the version provided by (Liu and Lapata, 2019a), which selects ranked top-40 paragraphs as input. For this dataset, we filter out documents whose summary length is less than 100 tokens. After the process, WikiSum test set contains 15,795 examples and the average length of summaries is 198.

WikiHow dataset is a large-scale dataset of instructions from the online WikiHow.com website (Koupaee and Wang, 2018). The task is to generate the concatenated summary-sentences from the paragraphs.

BillSum dataset contains US Congressional bills and human-written reference summaries from the 103rd-115th (1993-2018) sessions of Congress (Kornilova and Eidelman, 2019).

These datasets differ in scale, domain and task type. We collect details of the 8 corpus in Table 1.

4.2 Implementation Details and Metrics

FAR has 4 hyper-parameters and the best set of them are chosen from the following setting: $\alpha \in \{1, 2\}$, $\beta \in \{0.0, 0.1, \ldots, 0.9\}$, $\lambda_1 + \lambda_2 = 1$, $\lambda_1 \in \{0.0, 0.1, \ldots, 1.0\}$. In most case, FAR with the default setting ($\alpha = 1, \beta = 0.5, \lambda_1 = 0.5, \lambda_2 = 0.5$) can achieve satisfied performance on all datasets. We select best hyper-parameters by sampling 1,000

Datasets	Sources	Type	#Pairs			#Tokens	
			Train	Valid	Test	Doc.	Sum.
CNN/DM	News	SDS	287,227	13,368	11,490	788	63
NYT	News	SDS	36,735	5,531	4,375	1,291	80
MultiNews	News	MDS	44,972	5,622	5,622	2,104	264
arXiv	Scientific Paper	LDS	202,914	6,436	6,440	4,938	220
PubMed	Scientific Paper	LDS	117,108	6,631	6,658	3,016	203
WikiSum	Wikipedia	MDS	1,579,360	38,144	38,144	2,800	139
WikiHow	Wikipedia	SDS	157,252	5,599	5,577	581	63
BillSum	US Legislation	LDS	17,054	1,895	3,269	2,148	209

Table 1: Information of datasets. The data in Doc. and Sum. indicates the average length of document and summary respectively. SDS represents single-document summarization, MDS represents multi-documents summarization and LDS represents single long document summarization (#tokens of document $\geq$ 3,000).

Method	CNN/DM			NYT			WikiHow		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Oracle	52.59	17.62	36.67	61.63	41.54	58.11	39.80	14.85	36.90
PTR-GEN	39.50	17.30	36.40	**42.70**	**22.10**	**38.00**	-	-	-
REFRESH	41.30	18.40	35.70	41.30	22.00	37.80	-	-	-
BertExt	**43.25**	**20.24**	**39.63**	-	-	-	**30.31**	**8.17**	**28.24**
Lead	40.49	17.66	36.75	35.50	17.20	32.00	24.31	5.52	22.53
TextRank	33.85	13.61	30.14	33.24	14.74	29.92	21.64	5.34	19.68
LexRank	34.68	12.82	31.12	30.75	10.49	26.58	25.46	5.89	23.63
MMR	31.63	10.02	28.55	27.16	6.41	25.32	22.02	4.40	20.22
PacSum	40.70	17.80	36.90	41.40	21.70	37.50	-	-	-
PacSum (Ours)	40.69	17.82	36.91	41.37	21.65	37.35	27.46	6.13	25.40
STAS	**40.90**	**18.02**	**37.21**	41.46	21.80	37.57	-	-	-
FAR	40.83	17.85	36.91	**41.61**	**21.88**	**37.59**	**27.54**	**6.17**	**25.46**

Table 2: Results on SDS CNNDM, NYT and WikiHow test sets.

examples from validation set (Zheng and Lapata, 2019).

The implementation of our encoder model is based on the PyTorch implementation of BERT[*]. The BERT follows the base settings. In the post-training, we employ basic BERT model to initialize our sentence encoder. We use Adam (Kingma and Ba, 2014) as our optimizer with a learning-rate of $2e^{-5}$. During post-training, we sample documents from training set of all datasets. The max length of the input sentence is set to 60. A linear warm-up for the first 10% of steps followed by a linear decay to 0 is used. The BERT encoder is post-trained on 6 Tesla V100 GPUs.

We use `ROUGE-1.5.5.pl` script[†] to evaluated summarization quality automatically with ROUGE F1 (Lin and Hovy, 2003). We report ROUGE-1/2/L score to measure the quality of summaries. Besides, we also do a human evaluation for the facet bias and redundancy of extracted summaries.

4.3 Results

Table 2-4 report the results of datasets with 3 types. In each table, we present the results of **Oracle** and previous supervised models in the first block. **Oracle** can be seen as the upper bound of extractive models, which extracts gold standard summaries by greedily selecting sentences to optimize the mean of ROUGE-1 and ROUGE-2 (Nallapati et al., 2017). We compare our approach with strong unsupervised baselines **Lead**, **TextRank** (Mihalcea and Tarau, 2004), **LexRank** (Erkan and Radev, 2004), **MMR** (Carbonell and Goldstein, 1998) in the second block of each table. **Lead** selects the first k tokens as a summary. We also report previous best centrality-based model **PacSum** (Zheng and Lapata, 2019) in the third block of each table.

Overall, FAR outperforms above-mentioned unsupervised strong baselines on most datasets, especially on long-document and multi-documents datasets and is more generalized than them for differnt types, domains datasets.

Results on SDS Table 2 reports the results on single document summarization (SDS) datasets CNN/DM, NYT and WikiHow. **PTR-GEN** (See et al., 2017) is a supervised abstractive model with classic seq2seq structure. **REFRESH** (Narayan et al., 2018) and **BertExt** (Liu and Lapata, 2019b) are supervised extractive models. **STAS** (Xu et al., 2020) is the best unsupervised model on CNN/DM

[*]https://github.com/huggingface/transformers
[†]https://github.com/andersjo/pyrouge

Method	arXiv			PubMed			BillSum		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Oracle	53.88	23.05	34.9	55.05	27.48	38.66	56.22	38.77	51.25
PTR-GEN	32.06	9.04	25.16	35.86	10.22	29.69	**33.43**	**9.47**	**27.90**
Discourse-aware	35.80	11.05	31.8	38.93	15.37	35.21	-	-	-
SummaRuNNer	42.81	16.52	28.23	43.89	18.78	30.36	-	-	-
GlobalLocalCont	**43.62**	**17.36**	**29.14**	**44.85**	**19.70**	**31.43**	-	-	-
Lead	33.66	8.94	22.19	35.63	12.28	25.17	35.10	16.76	30.31
TextRank	24.38	10.57	22.18	38.66	**15.87**	34.53	36.10	15.00	30.35
LexRank	33.85	10.73	28.99	39.19	13.89	34.59	38.28	16.02	32.44
MMR	29.75	6.14	26.41	37.65	10.61	33.71	36.73	12.45	32.13
PacSum	39.33	12.19	34.18	39.79	14.00	36.09	38.34	16.64	33.36
FAR	**40.92**	**13.75**	**35.56**	**41.98**	15.66	**37.58**	**38.37**	**16.69**	**33.40**

Table 3: Results on LDS arXiv, PubMed and BillSum test sets.

Method	MultiNews			WikiSum		
	R-1	R-2	R-L	R-1	R-2	R-L
Oracle	55.40	29.91	50.51	49.43	27.18	45.04
FT (2019)	**44.32**	**15.11**	**20.50**	40.56	25.35	34.73
HT (2019)	42.36	15.27	22.08	**41.53**	**26.52**	**35.76**
T-DMCA (2018)	-	-	-	40.77	25.60	34.90
HiMAP (2019)	44.17	16.05	21.38	-	-	-
Lead	39.41	11.77	14.51	37.63	14.75	34.76
TextRank	38.44	13.10	13.50	23.66	7.79	21.23
LexRank	38.27	12.70	13.20	36.12	11.67	22.52
MMR	38.77	11.98	12.91	31.22	10.24	22.48
PacSum (2019)	43.27	14.16	38.25	36.85	12.94	33.64
FAR	**43.48**	**16.87**	**44.00**	**38.11**	**14.54**	**35.01**

Table 4: Results on MDS MultiNews and WikiSum test sets.

and NYT with two redesigned pretrain tasks to measure the importance of sentences.

From the results, we can see that: 1) Our model outperforms all strong baselines in the second block and PacSum by wide margins in terms of ROUGE-1/2/L on 3 SDS datasets. 2) Especially on NYT, our model outperforms the previous best unsupervised extractive system STAS and supervised method REFERSH.

After we re-implement the trigram blocking trick (i.e., removing sentences with repeating trigrams to existing summary sentences) which STAS used (Xu et al., 2020), FAR can achieve a better ROUGE-1 score **40.93**/17.80/37.00 than STAS on CNN/DM.

Results on LDS Table 3 reports the results on long document summarization (LDS) datasets arXiv, PubMed and BillSum. For supervised extractive models, we compare with **SummaRuNNer** (Nallapati et al., 2017) and **GlobalLocalCont** (Xiao and Carenini, 2019). We also compare with supervised abstractive models **Discourse-aware** (Cohan et al., 2018) and **PRT-GEN**.

As shown in Table 3, our model has obviously higher ROUGE-1/2/L score (+1.89 +1.56 +1.38) on arXiv and (+2.22 +1.55 +1.45) on PubMed than PacSum. Compare with supervised models, our unsupervised model outperforms supervised abstractive models PTR-GEN and Discourse-aware, but still have a gap with supervised extractive models. The reason for this gap is that supervised extractive models can extract sentences with dynamic length through training with labeled corpus, but unsupervised models need to predefined the length or number of extracted summaries.

Besides, we can see that the improvement on Billsum is limited. We analysis the input document of Billsum and find that documents in Billsum contains many very short sentences which lead to this limited improvement.

Results on MDS Table 4 reports the results on multi-documents summarization datasets MultiNews and WikiSum. **T-DMCA** and **HiMAP** are proposed with the construction of WikiSum and MultiNews. **FT** (Flat Transformer) and **HT** (Hierarchical Transformer) are two supervised extractive models which are proposed by (Liu and Lapata, 2019a).

From results in Table 4, we can see that PacSum and FAR have a strong performance on MultiNews, which may result from the characteristic of news datasets and the high-quality human-written documents-summary pairs of MultiNews. On Wik-

arXiv	
FAR	**40.92 13.75 35.56**
-facet-aware scoring	39.61 12.45 34.37
-modified DC	38.32 11.53 33.35
-post-training	40.02 12.79 34.67
NYT	
FAR	**41.67 21.93 37.68**
-facet-aware scoring	40.82 21.10 36.81
-modified DC	39.90 20.47 36.02
-post-training	40.93 21.38 36.99

Table 5: Ablation study on arXiv and NYT.

iSum, compare with PacSum, FAR is obviously better. We also can observe that the performance of unsupervised models are far less than supervised models. Because the length of multi-document summary has a great fluctuation and unsupervised methods are hard to decide the length of extracted sentences.

5 Analysis

In this section, we present a series of analysis and tests to understand the improvements of our FAR reported in the previous section, and to prove that it fulfills our intuition that the design of our model improves the facet bias. We choose NYT from SDS and arXiv from LDS to analyze the performance of FAR. These 2 datasets are typical and cover the situation of short and long document inputs.

Ablation Study In order to access the contribution of 3 components of FAR – modified DC in section 3.1, facet-aware scoring in section 3.2, and post-training in section 3.3. We remove each component of them and report ablation study results in 5. We can see that modified DC and facet-aware scoring are indispensable to the performance of FAR. If we remove each of them, the performance of FAR drops sharply. When we replace BERT with post-training with original BERT, the results also confirm that post-training is usable.

Human Evaluation To evaluate the ability of FAR in reducing facet bias and redundancy, we asked 3 human annotators to evaluate the extracted summaries of PacSum and FAR with the gold reference summary. Three annotators were asked to give 0-2 scores for facet bias and redundancy of 100 random sampled examples. The results of PacSum in terms of facet bias is 1.42 and redundancy is 1.17. Our FAR performs significantly better than PacSum ($p < 0.05$) whose facet bias is **0.96** and redundancy is **0.81**. Human evaluation results indicated that FAR can extract high-quality summaries by facet-aware modeling and reduce redundancy of summaries to some extent.

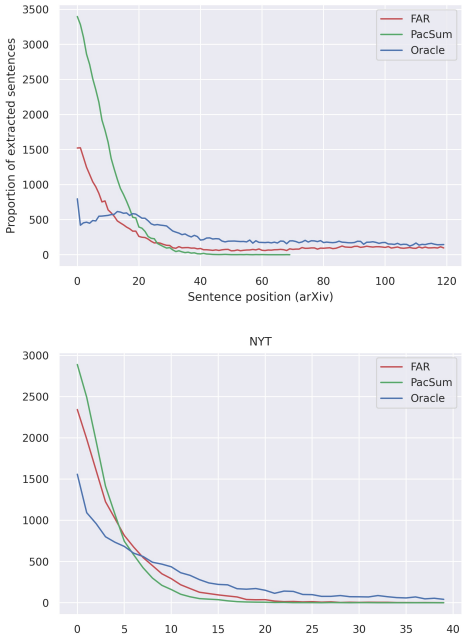

Figure 3: Sentence position distribution of arXiv and NYT. We use the first 40 sentences for NYT and the first 120 sentences for arXiv.

Sentence Position Distribution We compare the position distribution of extracted sentences of FAR, PacSum, and Oracle to further inspect the performance of FAR. We report the position distribution of extracted sentences in Figure 3. We can see that 1) The distribution of FAR is more close to Oracle; 2) PacSum only extracts sentence in the head of documents on arXiv, which is also mentioned by (Dong et al., 2020); 3) The advantages of our model are more significant for LDS datasets.

Analysis of Hyper-parameter β Hyper-parameter β is a crucial hyper-parameter that is used to filter out noise sentences in documents. We fixed other hyper-parameters and observed the change of ROUGE-1 from 0.1 to 0.9 with β in Figure 4. We can see that Hyper-parameter β has great impact on model's effect, especially on NYT dataset. These curves prove that noise sentences truly exists and hurt the performance of centrality-based models.

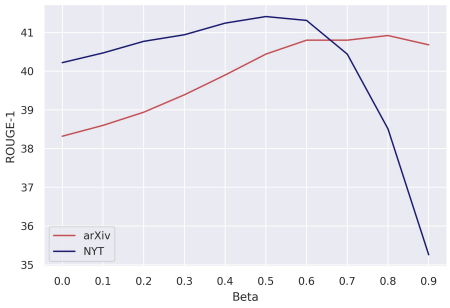

Figure 4: FAR's performance against different values of β on arXiv and NYT.

Case Study To intuitively show the ability of FAR to tackle the facet bias problem and reduce redundancy, we choose one typical example from NYT dataset. (*example is from a news report and only used to analyze the effectiveness of our model.*) As shown in Table 5, we can see that sentences extracted by PacSum all focus on the facet which describes terroristic attacks in Iraq. However, FAR can cover all 3 facets in gold reference. This shows that our FAR can effectively improve the performance by reducing the facet bias problem.

6 Related Work

Summarization is a long-standing challenge for researchers to address. Thanks to the power of the neural network and availability of large-scale parallel datasets. Supervised summarization algorithms develop sharply (Chopra et al., 2016; Cao et al., 2018; Zhang et al., 2018; Zhong et al., 2019; Gehrmann et al., 2019; Cho et al., 2019; Jin et al., 2020b; Cao et al., 2020; Jin et al., 2020a; Zhong et al., 2020). However, high-quality parallel datasets are not always available. Researches on unsupervised summarization are necessary, which can be diveided into extractive and abstractive. Unsupervised abstractive summarization is more challenging than extractive. There are also many interesting works (Wang and Lee, 2018; Févry and Phang, 2018; Baziotis et al., 2019; Jernite, 2019; Zhou and Rush, 2019; West et al., 2019; Chu and Liu, 2019; Yang et al., 2020) on unsupervised abstractive summarization.

However, most unsupervised summarization models are extractive (Radev et al., 2000; Mihalcea and Tarau, 2004; Erkan and Radev, 2004; Carbonell and Goldstein, 1998; Wan, 2008; Wan and Yang, 2008; Schluter and Søgaard, 2015; Zhao et al., 2020) and focused on the measure of sentence salient. Graph-based models are effective and widely concerned in unsupervised extractive methods. Different from traditional undirected graph rank models (Radev et al., 2000; Mihalcea and Tarau, 2004; Erkan and Radev, 2004), (Zheng and Lapata, 2019) proposed directed centrality method, which is based on the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) assumption. (Dong et al., 2020) point out that PACSUM has position bias, which makes PACSUM not suitable for long document summarization, and proposed hierarchical position-based model HipoRankfor scientific document summarization. STAS (Xu et al., 2020) design two summarization tasks related pretraining tasks to improve sentence representation. Then they proposed a rank method which combines attention weight with reconstruction loss to measure the centrality of sentences.

We find the facet bias problem in graph-based models, which lead to the extracted summaries can not cover multi-facets information in document. A similar concept in summarization is redundancy. However, the difference between redundancy and facet bias is two folds: 1) to solve redundant problem, we just need to make sure selected sentences are not too similar; 2) However, to tackle the facet bias problem, we need to select sentences that can retain multi-facets information.

7 Conclusion

In this paper, we discover the facet bias problem in centrality-based unsupervised summarization models and proposed a novel facet-aware centrality-based ranking model FAR to tackle it. We introduce a sentence-document weight into centrality, which forced the model to pay more attention to different facets and find that FAR can reduce redundancy to some extent. Results on a wide range of summarization tasks show that our method consistently outperforms strong baselines especially in long- and multi-document scenarios, which prove our model is robust and effective. Extensive analyses confirmed that the performance gains of our model come from alleviating the facet bias problem.

Acknowledgments

We thank the three anonymous reviewers for their careful reading of our paper and their many insightful comments and suggestions. This work was supported in part by the National

Gold Reference
1. Marine corps lt col jeffrey r. chessani was battalion commander in haditha, iraq, when roadside bomb planted by sunni arab insurgents killed one of his marines and wounded two other, and when infantrymen under his command, seeking to engage enemy, instead killed 24 civilians.
2. Chessani and other officers are charged with dereliction of duty for failing to investigate episode properly and relieved of their command.
3. Military hearing at camp pendleton will decide if chessani should face formal court-martial. Witnesses and military documents paint two contradictory portraits of chessani, highest-ranking marine officer charged since iraq war began.

PacSum
1. In november 2005, colonel chessani was a battalion commander in haditha, iraq, when a roadside bomb planted by sunni arab insurgents killed one of his marines and wounded two others.
2. Through three combat deployments in iraq, a bronze star and numerous combat ribbons, lt col jeffrey r. chessani's marine corps career has been defined, it seems, by terrorist bombs.
3. In october 1983, news of the attack by muslim extremists on a marine corps barracks in beirut that killed 241 service members compelled colonel chessani, then a teenager from rangely, colo., to embrace christianity and, later, to follow two brothers into the service.

FAR
1. In november 2005, colonel chessani was a battalion commander in haditha, iraq, when a roadside bomb planted by sunni arab insurgents killed one of his marines and wounded two others.
2. Last year, the marine corps charged colonel chessani, 43, and three other officers with dereliction of duty for failing to investigate the episode properly and relieved him of his command.
3. In a military hearing into whether he should face a formal court-martial, witnesses and military documents have helped paint two contradictory portraits of colonel chessani, the highest-ranking marine officer charged since the iraq war began more than four years ago.

Figure 5: The examples come from New York Times dataset.

Natural Science Foundation of China (Grant Nos.U1636211, 61672081, 61370126), the 2020 Tencet Wechat Rhino-Bird Focused Research Program, and the Fund of the State Key Laboratory of Software Development Environment (Grant No.SKLSDE2019ZX-17).

References

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. SEQˆ3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jaime Carbonell and Jade Goldstein. 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*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

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. Association for Computational Linguistics.

Sangwoo Cho, Logan Lebanoff, Hassan Foroosh, and Fei Liu. 2019. Improving the similarity measure of determinantal point processes for extractive multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1027–1038, Florence, Italy. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. MeanSum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232, Long Beach, California, USA. PMLR.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621,

New Orleans, Louisiana. Association for Computational Linguistics.

Yue Dong, Andrei Romascanu, and Jackie CK Cheung. 2020. Hiporank: Incorporating hierarchical and positional information into graph-based unsupervised long document extractive summarization. *arXiv preprint arXiv:2005.00513*.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 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)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

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

Thibault Févry and Jason Phang. 2018. Unsupervised sentence compression using denoising auto-encoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Gehrmann, Zachary Ziegler, and Alexander Rush. 2019. Generating abstractive summaries with finetuned language models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 516–522, Tokyo, Japan. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pages 1693–1701. Curran Associates, Inc.

Yacine Jernite. 2019. Unsupervised text summarization via mixed model back-translation. *arXiv preprint arXiv:1908.08566*.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020a. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6244–6254, Online. Association for Computational Linguistics.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020b. Semsum: Semantic dependency guided neural abstractive summarization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8026–8033. AAAI Press.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

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

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences.

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

Yang Liu and Mirella Lapata. 2019b. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

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

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 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)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2011. *Automatic summarization*. Now Publishers Inc.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Natalie Schluter and Anders Søgaard. 2015. Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 840–844, Beijing, China. Association for Computational Linguistics.

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

Antoine Tixier, Polykarpos Meladianos, and Michalis Vazirgiannis. 2017. Combining graph degeneracy and submodularity for unsupervised extractive summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 48–58, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaojun Wan. 2008. An exploration of document impact on graph-based multi-document summarization. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 755–762, Honolulu, Hawaii. Association for Computational Linguistics.

Xiaojun Wan and Jianwu Yang. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 299–306, New York, NY, USA. Association for Computing Machinery.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing 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, Online. Association for Computational Linguistics.

Yaushian Wang and Hung-Yi Lee. 2018. Learning to encode text as human-readable summaries using generative adversarial networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4187–4195, Brussels, Belgium. Association for Computational Linguistics.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3752–3761, Hong Kong, China. Association for Computational Linguistics.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3011–3021, Hong Kong, China. Association for Computational Linguistics.

Shusheng Xu, Xingxing Zhang, Yi Wu, Furu Wei, and Ming Zhou. 2020. Unsupervised extractive summarization by pre-training hierarchical transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1784–1795, Online. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874, Online. Association for Computational Linguistics.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference*

on Empirical Methods in Natural Language Processing, pages 779–784, Brussels, Belgium. Association for Computational Linguistics.

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

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. SIGIR '20, page 1949–1952, New York, NY, USA. Association for Computing Machinery.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. 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, Florence, Italy. Association for Computational Linguistics.

Jiawei Zhou and Alexander Rush. 2019. Simple unsupervised summarization by contextual matching. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5101–5106, Florence, Italy. Association for Computational Linguistics.

A Hyperparamters

Hyper-paramters of the FAR were reported in Table 6.

B Filter Summary Length of arXiv&PubMed

To prove unsupervised is limited by summary length, we filter examples in the test set with summary length, and report the results in Figure 6. We can see that when examples with short summary, which do not match the predefined length, were removed, the performance improved obviously.

Datasets	α	β	λ_1	λ_2
CNN/DM	1	0.0	0.7	0.3
NYT	1	0.6	0.6	0.4
arXiv	2	0.7	0.5	0.5
PubMed	2	0.3	0.5	0.5
MultiNews	1	0.4	0.5	0.5
WikiSum	1	0.0	0.5	0.5
BillSum	1	0.5	0.5	0.5
WikiHow	1	0.8	0.5	0.5

Table 6: Hyper-parameters for FAR's best performance.

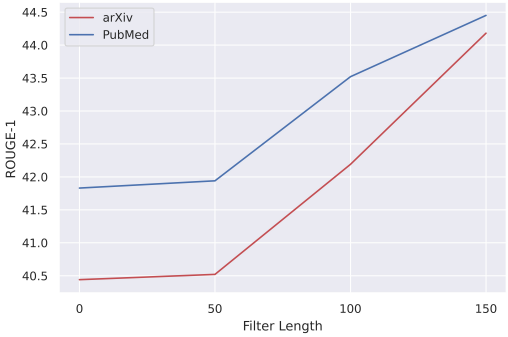

Figure 6: Performance on arXiv and PubMed, when we filter examples in test set with summary lenghth.

C Sentence Position Distribution

We show sentence position distribution of all 8 datasets in Figure 7.

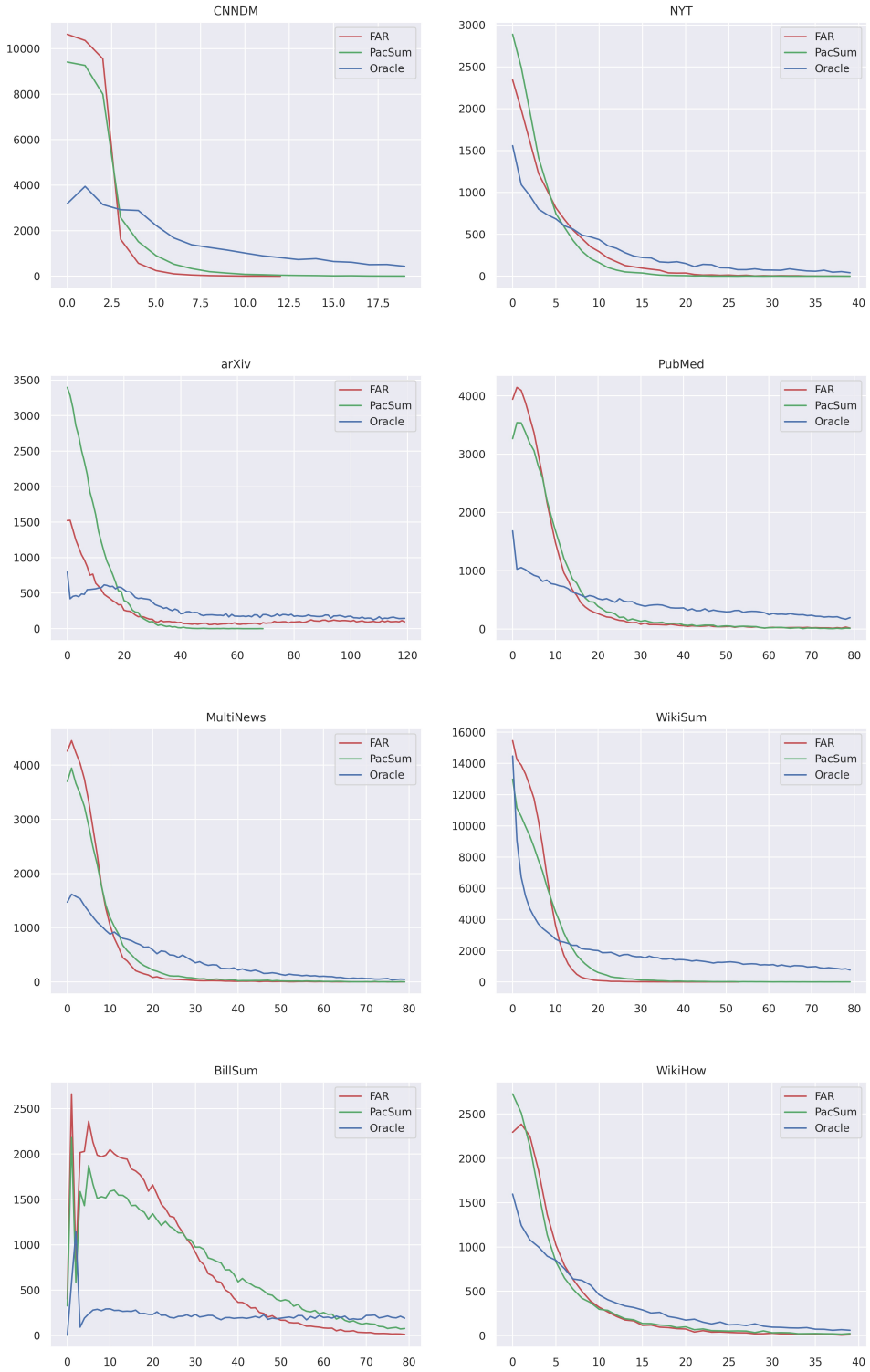

Figure 7: Sentence position distribution of 8 datasets.