FactoScalpel: Enhancing the Factual Consistency of Abstractive Summarization through Knowledge Injection

Anonymous ACL submission

Abstract

Recently, abstractive summarization has problems with factual inconsistencies in generated summaries. Inspired by the related work on knowledge storage in Transformer, we firstly explore the relationship between factual errors and Feed-Forward Networks (FFNs) in Transformer, and propose factual errors attribution method. Based on the results, we inject knowledge to the decoder for the first time, propose a fact-aware summarization model FactoScalpel which integrates a Knowledge Bank and router-controlled mechanism into FFNs. By introducing facts through Knowledge Bank, balancing the original FNN with the newly added Knowledge Bank module through router-controlled mechanism, FactoScalpel achieves factual improvement of the decoder end through fine surgery. We compare FactoScalpel with multiple fact-aware summarization models using multiple factual consistency metrics based on the XSum, our method achieves state-of-the-art results in most experiments.

1 Introduction

Recently, the issue of factual inconsistencies has continued to be a focus for researchers. Both post-editing methods (Chen et al., 2021; Lee et al., 2022; Li et al., 2024) and end-to-end summarization models demonstrate their effectiveness. However, the performance of the former is inherently dependent on the quality of the model-generated summaries. As a result, we focus more on latter improvements.

End-to-end summarization models improve the summarization model directly, which are also called fact-aware summarization models. As shown in Figure 1, fact-aware summarization models are divided into four categories, metric-based, decoding-based, multitasking-based and knowledge injection-based methods.

Both metric-based (Cao et al., 2021; Nan et al., 2021; Wu et al., 2022) and decoding-based methods (van der Poel et al., 2022; King et al., 2022a; Sridhar and Visser, 2022) employ factual consistency metrics to improve the factual consistency in model-generated summaries. However, the former is used as an optimization targets, while the latter introduces additional factual scores to the decoding process. On the other hand, multitasking-based methods (Cao and Wang, 2021a; Tang et al., 2021; Zhang et al., 2022; Wang et al., 2022a) introduce additional tasks to improve the summarization models, including contrastive learning and control codes. Furthermore, knowledge injection-based methods (Zhu et al., 2020; Dou et al., 2020; Dong et al., 2022b; Xu and Zhao, 2022) offer superior interpretability compared to the aforementioned approaches. By explicitly defining the facts in documents and integrating them into the summarization models, these methods bolster the model's awareness of factual information.

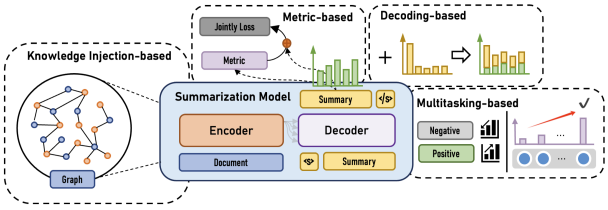

Figure 1: Fact-aware summarization models based on different methods.

We consider knowledge injection-based methods as they hold the potential to serve as a foundational framework that could enhance other methodologies. Most of recent studies focus on incorporating facts into the encoder embeddings or final outputs through a factual encoder. However, there is a lack of exploration into the internal workings of the decoder, which is a critical component for generating summaries.

Our contributions are summarized as follows:

- Inspired by related studies on knowledge storage of FFNs in Transformer, we firstly study the relationship between the FFNs and fac-

tual errors and propose factual errors attribution method. We examine the decoder's internal mechanisms, and firstly explore the fine-grained relationship between the decoder and hallucinations.

- Based on the attribution analysis, we propose a fact-aware summarization model FactoScalpel[1], which is based on BART and adds Knowledge Bank and router-controlled mechanism on the decoder side. Knowledge Bank contains factual factors from document, which are integrated into the FFN of the decoder. Router-controlled mechanism reduces the "exclusion reaction" caused by introducing the Knowledge Bank.

- Through a series of comprehensive experiments conducted on the XSum dataset, FactoScalpel consistently outperforms competitive models FASum, CLIFF, PINOCCHIO and CoFE in most of the fact consistency metrics QAFactEval, SummaC, ClozE, DAE and FactCC, demonstrating the excellent efficacy of our method in maintaining factual consistency.

2 Related Work

2.1 Fact-aware Summarization Models

Metric-based Cao et al. (2021) and Nan et al. (2021) directly use automatic evaluation metric as part of the loss function. Wu et al. (2022) proposes a fact robustness evaluation metric based on the probability of generating correct and incorrect factual fragments in the summary, and propose a training strategy based on adversarial learning.

Decoding-based van der Poel et al. (2022) proposes a decoding strategy when the entropy of the generated text exceeds a certain threshold, the logarithmic probability will be subtracted from a logarithmic marginal probability. Sridhar and Visser (2022) directly uses factual consistency evaluation indicators as the basis for decoding optimization, and consider both the probability and the factual scores given by the indicators when selecting words.

Multitasking-based Cao and Wang (2021a) and Tang et al. (2021) introduce contrastive learning and construct positive and negative samples to train the model. Zhang et al. (2022) and Wang et al. (2022a) divide training data into subsets based on factuality, and each subset is assigned a control code. They select the code with the highest factuality by control code to generate high factual consistency summary.

Knowledge Injection-based Zhu et al. (2020) extracts and integrates factual relations into the summary generation process via graph attention. Dong et al. (2022b) introduces the knowledge graph, serializing the subgraphs of the knowledge graph and inputting them into the encoder together with the original text to generate summaries.

2.2 Knowledge Storage

Knowledge in FFNs Petroni et al. (2019) reveals that pre-trained language models (PLMs) contain knowledge competitive with traditional NLP methods. Furthermore, Geva et al. (2021) shows that FFNs in transformer-based language models could be considered as key-value memories, where each key correlates with textual patterns in the training examples, and each value induces a distribution over the output vocabulary. Based on these insight, Dai et al. (2022) introduces a novel knowledge attribution algorithm and improve the performance of PLMs by directly editing knowledge neurons.

Knowledge Injection Based on prior work, numerous studies improve the PLMs by enhancing the knowledge capability of FFNs. Dong et al. (2022a) first detects the fake facts stored in FFNs and introduce lightweight parameters to calibrate them. Similarly, Yao et al. (2022) integrates lightweight parameters to the FFNs and injects external knowledge to them. Dai et al. (2023) designs a neural knowledge bank module and a knowledge injection strategy to introduce external knowledge for PLMs via pre-training. However, the purpose of these studies is to expand the knowledge base of PLMs. In contrast, our work aims to ensure that the factual knowledge within PLMs remains aligned with the content of the corresponding documents.

3 Factual Errors Attribution

In order to explore the relationship between FFNs and factual errors generated by summarization models, we introduce and extend the attribution algorithm proposed by Dai et al. (2022). Then, we will analyze and summarize the attribution results.

[1]The newly added Knowledge Bank and router-controlled mechanism are the same as organ transplantation and neural connections, we use these for facts surgery.

3.1 Factual Attribution Algorithm

The attribution algorithm proposed by Dai et al. (2022) is primarily designed for encoder-only models, and the investigation into factual units is limited to single-token instances. In our work, we explore the factual errors which define factual factors [2] and consist of multiple tokens. Meanwhile, the factual errors are generated by the auto-regressive decoder. Therefore, we modified the formula for calculating the attribution score in Dai et al. (2022), as shown in Appendix A.1.

Next, we set a threshold to filter out neurons with low scores. After that, each token in one factual error will respond to several neurons. In order to refine the neurons, we set another threshold to filter out the neurons which do not co-occur in most of tokens. The retained neurons are considered as the knowledge neurons to the corresponding factual errors. More detailed description is shown in Appendix A.2.

3.2 Factual Errors Attribution Analysis

To ensure our study is in step with prevalent summarization models, we select BART (Lewis et al., 2020) trained on the XSum (Narayan et al., 2018) as our research object. We conduct an analysis of knowledge neuron distributions by randomly selecting a subset of 400 samples for examination.

3.2.1 Distribution of Knowledge Neurons

Figure 2 illustrates the distribution of knowledge neurons in each BART layer. The left side of the vertical axis represents the ratio of knowledge neurons per layer, while the right side denotes the average quantity of knowledge neurons. We note that knowledge neurons tend to be concentrated in the high-level FFNs. Meanwhile, the absolute number of knowledge neurons remains relatively low. Both of these facts demonstrate the relevance and concentration of factual errors with FFNs.

3.2.2 Hallucinations and Knowledge Neurons

To explore the relationship between hallucinations and knowledge neurons, we first propose a simple metric to measure the degree of instrinsic and extrinsic hallucinations, as follows:

$$IEScore = \max_{w \in f \cap w \in X} \frac{|w|}{|f|} \quad (1)$$

where w represents a continuous token fragment among the factual factors f in the summary, and

[2] Factual factors indicate as entities and noun phrases.

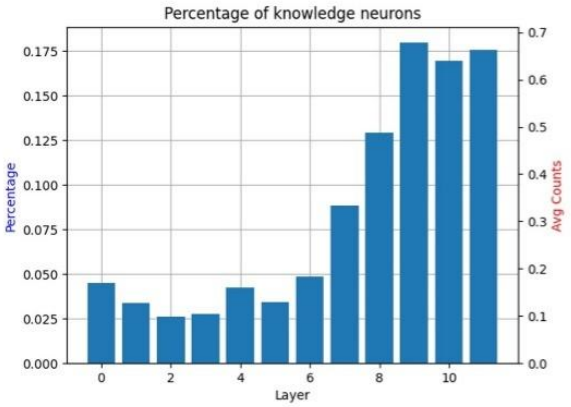

Figure 2: Distribution of identified knowledge neurons in each BART layer.

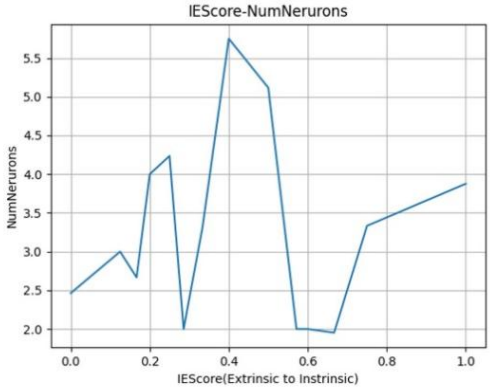

Figure 3: Correlation between knowledge neurons and hallucinations in BART.

$\boldsymbol{X}$ represents the document corresponding to the summary. A score approaching 0 suggests that the factual factor f is likely an extrinsic hallucination, otherwise it could be an intrinsic hallucination.

As shown in Figure 3, there is a notable correlation between FFNs and the generation of both intrinsic and extrinsic hallucinations. This suggests that by enhancing the architecture of FFNs, we may concurrently mitigate the occurrence of both types of hallucinations.

4 FactoScalpel

The analysis in section §3.2 establishes a link between FFNs and the factual errors. Building on this insight, we propose a novel knowledge injection-based method FactoScalpel, which is a promising approach to augment factual consistency in summaries by integrating factual factors from the document into the FFNs of the decoder.

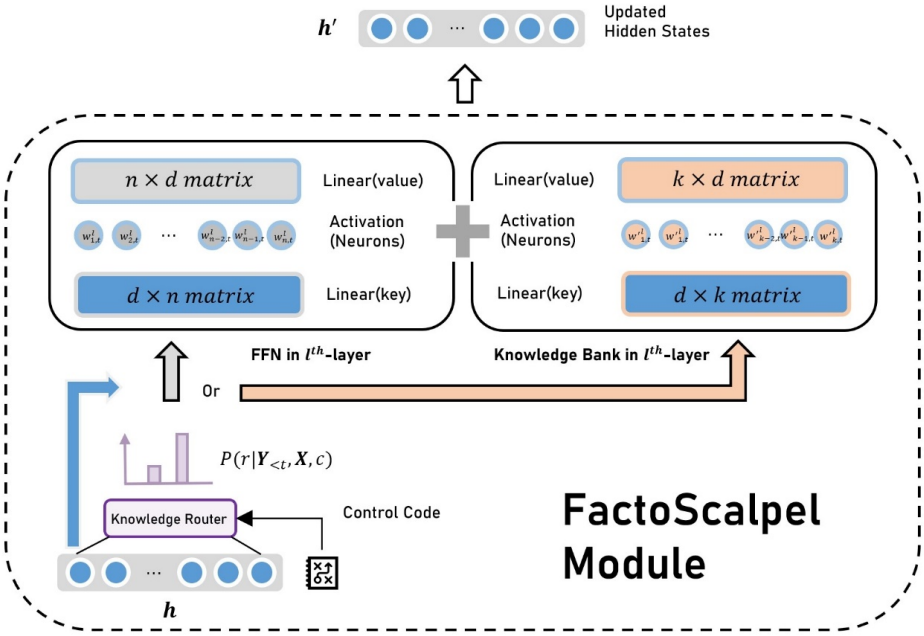

Figure 4: Overview of a FactoScalpel module.

4.1 Knowledge Bank

As shown in Figure 4, we define $\boldsymbol{h} \in \mathbb{R}^{1\times d}$ as the hidden state to the i-th FFN, and the updating process of the vanilla FFN is as follows:

$$\boldsymbol{h}' = \sigma\left(\boldsymbol{h}\cdot\boldsymbol{W}_1^\top\right)\cdot\boldsymbol{W}_2 \tag{2}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{n\times d}$ are learnable parameters. d and n represent the dimension of the vectors and the number of neurons. $\sigma(*)$ denotes the any activation function in FFNs. We ignore the bias in FFNs in convenience.

According to Geva et al. (2021), $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ can be considered as key-value memory. Therefore, we can build a similar FFN structure $\boldsymbol{W}_1^{(KB)}$ and $\boldsymbol{W}_2^{(KB)}$ to extend the original key-value memory to inject factual knowledge into the decoder, which is called **Knowledge Bank**. To generate the embeddings for the factual factors $h^{(f)}$, we compute them by taking the mean of the word embeddings for each token.

Then, we introduce a pair of learnable mapping matrices $\boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d\times d}$ to connect the feature space between Knowledge Bank and FFN. For the embeddings of factual factors $\boldsymbol{h}^{(f)}$, a Knowledge Bank can be formulated as follows:

$$[\boldsymbol{W}_1^{(KB)}, \boldsymbol{W}_2^{(KB)}] = h^{(f)}\cdot[\boldsymbol{W}_k, \boldsymbol{W}_v] \tag{3}$$

Composed with Knowledge Bank, the process of updating hidden states $\boldsymbol{h}$ can be reformulated as follows:

$$\begin{aligned}\boldsymbol{h}' = &\sigma\left(\boldsymbol{h}\cdot\boldsymbol{W}_1^\top\right)\cdot\boldsymbol{W}_2 \\ &+\sigma\left(\boldsymbol{h}\cdot\left(\boldsymbol{W}_1^{(KB)}\right)^\top\right)\cdot\boldsymbol{W}_2^{(KB)}\end{aligned} \tag{4}$$

Eq.4 significantly enhances the capacity of the FFNs, which explicitly integrate factual knowledge from documents into the forward propagation process of the decoder.

4.2 Knowledge Router

According to the key-value memory theory for FFNs, we expect that the incorporating Knowledge Banks into FFNs should enhance the ability of the summarization model to reference pertinent information within a document, thereby bolstering the factual consistency of generated summaries. Despite this potential improvement, the current approach suffers from limited interpretability and lacks control mechanisms. To address these shortcomings, we need to establish a more robust structure for formulating specifications for module selection within the summarization process.

Factually, the Knowledge Bank can be conceptualized as a Mixture of Experts (MoE) system comprising two experts. Consequently, we adopt the MoE's "Router" mechanism to develop our Knowledge Router module. This module effectively generates a probability distribution over the selection of the respective experts. In this study, the

Knowledge Router is designed to allocate weights to the two components, mitigating the "rejection response" that may arise from integrating the Knowledge Bank.

Similarly, we define $\boldsymbol{h} \in \mathbb{R}^{1\times d}$ as the hidden state to the i-th FFN and introduce a learnable mapping matrix $\boldsymbol{W}_R \in \mathbb{R}^{d\times 2}$ to calculate the weight distribution as follows:

$$P\left(r \mid \boldsymbol{Y}_{<t}, \boldsymbol{X}\right) = \operatorname{softmax}\left(\boldsymbol{h} \cdot \boldsymbol{W}_R\right) \quad (5)$$

where t represents the current prediction step of decoder and r denotes the routes to a FFN and Knowledge Bank module. $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote a document and a generated summary separately. We define the probability p as the likelihood of selecting the Knowledge Bank given the previously generated outputs $\boldsymbol{Y}_{<t}$ and the input sequence $\boldsymbol{X}$. Consequently, this allows us to reformulate the updating process presented in Eq.6 to incorporate this probabilistic decision-making process.

$$\begin{aligned}\boldsymbol{h}' = (1-p)\cdot\sigma\left(\boldsymbol{h}\cdot\boldsymbol{W}_1^{\top}\right)\cdot\boldsymbol{W}_2 + \\ p\cdot\sigma\left(\boldsymbol{h}\cdot\left(\boldsymbol{W}_1^{(KB)}\right)^{\top}\right)\cdot\boldsymbol{W}_2^{(KB)}\end{aligned} \quad (6)$$

Similar to the expert selection challenge in MoE, FactoScalpel exhibits a bias towards selecting the "FFNs pre-trained with extensive data" throughout the finetuning phase, consequently neglecting the randomly initialized Knowledge Banks[3]. To mitigate this, we introduce a router loss to balance the selection probabilities between the FFNs and the Knowledge Banks.

We suppose that the summarization model should engage the route of Knowledge Banks when it synthesizes factual factors that are explicitly present in the document. To facilitate this, we propose a sequence tagging task for Knowledge Router with two labels. For each factual factors within the summary Y, if it is completely appeared in the source document X, we assign a label $L^{(R)}$ of 1 to every token in the factual factor; otherwise, the label if each token will be set to 0. Consequently, for each document-summary pair, we define the router loss as follows:

$$\begin{aligned}\mathcal{L}_R = -\frac{1}{2|\boldsymbol{Y}|}\sum_{t=1}^{|\boldsymbol{Y}|}[L_t^{(R)}\cdot\log p_t \\ + \left(1-L_t^{(R)}\right)\cdot\log\left(1-p_t\right)]\end{aligned} \quad (7)$$

where t denotes the current generation step of the decoder.

4.3 Control Codes

Recent work (Nan et al., 2021; Guo et al., 2022; Li et al., 2024) indicates that summarization datasets also contain numerous factual errors. The mainstream solution is filtering the unfaithful samples with factual consistency metrics. However, it is difficult to choose the exact filtering threshold. High threshold will discard most samples, resulting as lack of samples (Adams et al., 2022). To balance the faithfulness and the utility of the available data, we leverage the control code mechanism (Zhang et al., 2022; Wang et al., 2022a) on filtering datasets SummDSC-base (Li et al., 2024).

Specially, we introduce control codes into Knowledge Router through vector embedding instead of prompting tokens. $\boldsymbol{W}_c \in \mathbb{R}^{n_c\times d}$ is defined as a learned matrix, where n_c is the number of control codes. For a summary and its corresponding control code c, Eq.5 will be updated as follows:

$$\begin{aligned}P(r \mid \boldsymbol{Y}_{<t}, \boldsymbol{X}, \boldsymbol{c}) \\ = \operatorname{softmax}\left(\left(\boldsymbol{h} + \boldsymbol{W}_c[c,:]\right)\cdot\boldsymbol{W}_R\right)\end{aligned} \quad (8)$$

where $\boldsymbol{W}_c[c,:]$ means taking the c-th row vector.

Due to the router loss, we use $IEScore$ to divide the summaries in the training dataset into three levels, **Low**, **Medium** and **High**.

4.4 Training & Inference

As discussed in §3.2, the high-level layers of the decoder are more relative with factual factors. Therefore, we place the FactoScalpel module to the final layer to refine and enhance the output with minimal disruption to the overall architecture. Meanwhile, to bridge the gap between the newly integrated knowledge module and the existing parameters of the pre-trained model, we have structured the training process into pre-training and fine-tuning.

During the pre-training phase, we incorporate the Knowledge Bank module while keeping all other model parameters frozen. The pre-training dataset is composed of documents from the summarization datsets. We use a masked language

[3]Since $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ are randomly initialized, the Knowledge Bank will be chaotic even if $\boldsymbol{h}^{(f)}$ is deterministic according to Eq.3.

Model	QAFactEval	SummaC	ClozE	DAE	FactCC	R-1	R-2	R-L
BART-large	18.48	9.06	69.97	61.20	22.69	43.68	20.25	35.22
FASum	3.63	5.97	56.16	43.01	27.10	30.31	10.02	23.76
CLIFF	14.18	9.48	68.52	58.47	25.24	40.21	17.66	32.40
PINOCCHIO	0.07	0.86	39.61	42.61	**31.50**	42.69	19.16	33.85
CoFE	20.41	11.24	73.94	64.86	24.87	**44.90**	**21.90**	**36.75**
FactoScalpel (ours)	**20.63**	**13.45**	**75.93**	**69.35**	28.13	40.92	17.48	32.74

Table 1: Results of factual consistency metrics and ROUGE for each summarization model based on the XSum testset. The highest performance across these metrics is highlighted in bold.

modeling task for pre-training, where the masking is applied to factual factors within the documents to align with the Knowledge Bank.

During fine-tuning phase, we introduce knowledge router and control code to build a complete FactoScalpel. The optimization objectives include cross-entropy loss $\mathcal{L}_{CE}$ and router loss $\mathcal{L}_R$ of the summary. The combined loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_R \tag{9}$$

where λ is a hyperparameter to control the impact of knowledge router.

During inference, we only need to input the document into the encoder and set the control code to **High** to generate the faithful summary.

5 Experiments

5.1 Experimental Settings

Benchmark and Evaluation metrics We train FactoScalpel on SummDSC-base (XSum) and conduct the experiments on raw XSum testset. We use ROUGE-1, ROUGE-2 and ROUGE-L to evaluate the informativeness while five factual consistency metrics to evaluate factual consistency, inculding QAFactEval(Fabbri et al., 2022), SummaC(Laban et al., 2022), ClozE(Li et al., 2023), DAE(Goyal and Durrett, 2020) and FactCC(Kryscinski et al., 2020). Moreover, we also evaluate these models through human evaluation.

Baselines We select five open source work as our baselines, including FASum(Zhu et al., 2021), CLIFF(Cao and Wang, 2021b), PINOCCHIO(King et al., 2022b) and CoFE(Wang et al., 2022b). Among these baselines, FASum is knowledge injection-based methods, which is same to FactoSaclpel. CLIFF and CoFE are multitasking-based methods and utilize contrastive learning. PINOCCHIO is a decoding-based methods and generates summaries with a factual consistency metric. Moreover, we have also incorporated the basic summarization model BART trained on XSum.

5.2 Implementation Details

Following the fact-aware summarization models in baselines, we also choose BART as the basic skeleton for FactoScalpel to ensure a fair and credible comparison of model performance. We use the `en_core_web_trf` model from SpaCy [4] to extract factual factors and use BART-large in Huggingface [5] to initialize the parameters of our models. To pre-train the Knowledge Bank in FactoScalpel, we adopt the AdaFactor(Shazeer and Stern, 2018) optimizer with learning rate 1e-4 to train for 10 epochs with batch size 32 on two NVIDIA GeForce RTX 3090 GPUs. During fine-tuning, we use the AdamW optimizer with learning rate 1e-5 to train for 5 epochs with batch size 8 on the same computing resource configuration.

5.3 Performance on XSum Benchmark

As shown in Table 1, we show the results of factual consistency metrics and ROUGE for each summarization model based on the XSum testset. FactoScalpel achieves the best results on most factual consistency metrics. However, there is a slight degree of decline in ROUGE. We believe it is an expected outcome, as ROUGE metric prioritizes similarity over factual consistency. Because golden summaries in XSum have certain factual errors, it must lead to difference with the golden summaries when the fact-aware summarization model generates summaries without these factual errors.

Compared to other fact-aware summarization models, FASum only has good results in FactCC. Meanwhile, its performance on ROUGE significantly diminishes, signaling a marked reduction in

[4] https://spacy.io/api
[5] https://https://huggingface.co

Model	Info.	Con.	In. H	Ex. H
BART-large	3.31	2.86	0.56	0.81
FASum	2.33	1.89	0.85	1.65
CLIFF	**3.59**	2.76	0.95	1.04
PINOCCHIO	3.43	2.89	0.94	0.90
CoFE	3.50	3.30	0.44	1.05
FactoScalpel	3.56	**3.53**	**0.41**	**0.87**

Table 2: Human evaluation results of different fact-aware summarization models.

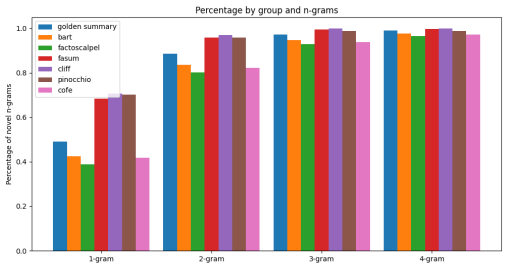

Figure 5: Percentage of novel n-grams for summaries

its summarization capabilities. This trend implies that the knowledge injection approach employed by FASum not only undermines its ability to generate summaries but also fails to enhance its capacity for factual consistency.

For the multitasking-based methods, CoFE performs better than CLIFF and has certain improvements in several factual consistency metrics. At the same time, CoFE achieves the best results on ROUGE, which shows that its generated summaries are more similar to golden summaries. It may also be one of the reasons why its factual consistency is lower than those on FactoScalpel.

In the case of PINOCCHIO, while it excels in achieving top results on FactCC, it simultaneously exhibits notably weak performance across other factual consistency metrics. This discrepancy could stem from an excessive dependence on a solitary evaluation criterion during the decoding process.

5.4 Human Evaluation

We manually evaluate 90 samples randomly selected from the XSum testset, labeling the informativeness (Info.), consistency (Con.), the number of instrinsic hallucination (In. H) and the number of extrinsic hallucination (Ex. H). We divide the 90 samples into two groups, namely Group A with 80 samples and Group B with 10 samples. For Group A, 80 samples are assigned equally to all annotators. And the samples in Group B will be assigned to all annotators. We use the results in Group B to calculate the consistency between different annotators to ensure the credit of annotation results.

The results of the human evaluation are presented in Table 2, where the average Cohen's Kappa score is 0.615. It is worth noting that FASum exhibits subpar performance, falling short of the basic BART in terms of informativess and factual consistency. The performance of PINOCCHIO is better than that in §5.3 and ranks in the middle tier of summarization models alongside CLIFF, both showing modest enhancements over the basic BART. At the forefront, CoFE and FactoScalpel stand out, delivering high-quality summaries characterized by strong factual accuracy and minimal errors, without compromising on summarization capabilities. Moreover, FactoScalpel outperforms all other models in our comparison.

5.5 Novel N-grams

It is important to examine the number of novel grams in model-generated summaries. Previous work indicates that high-quality summaries typically maintain a proportion of novel words similar to golden summaries in abstractive summarization. However, the faithulness of summarization datasets is often overlooked. As discussed in §5.3, if golden summaries contain numerous factual errors, it may be reasonable and acceptable for a model to generate fewer new words in comparison. This is because accurately representing a fact usually involves reusing words from the document, which naturally leads to a lower rate of novel word introduction. Nonetheless, this reduction should remain within a reasonable threshold to ensure quality.

In order to measure the proportion of novel words, we count the ratio of novel 1-gram, 2-gram, 3-gram and 4-gram in summaries. The results are shown in Figure 5. FASum, CLIFF and PINOCCHIO all show a higher proportion of novel grams than the golden summaries and BART. However, comparing the factual consistency evaluation results in Table 1 and 2, this may also introduce more extrinsic hallucinations. For CoFE and FactoScalpel, their proportion of novel grams is lower than that of the golden summaries and BART, but it is still within an acceptable range and has not declined to the category of extractive summarization.

Model	QAFactEval	SummaC	ClozE	DAE	FactCC	R-1	R-2	R-L
FactoScalpel	20.63	**13.45**	**75.93**	**69.35**	**28.13**	40.92	17.48	32.74
-SummDSC-base	18.88	11.20	73.66	64.76	25.47	42.75	19.24	34.66
-Control Code	18.37	10.12	73.22	64.30	24.49	42.54	18.93	34.17
-Knowledge Router	18.60	9.45	73.17	64.42	23.71	43.16	19.43	34.59
-Pre-training	18.59	9.29	72.93	64.38	23.67	43.18	19.53	34.64
-Knowledge Bank	18.48	9.06	69.97	61.20	22.69	**43.68**	**20.25**	**35.22**
BART + SummDSC-base	**20.75**	11.90	73.70	67.63	26.43	40.30	16.91	31.98

Table 3: Results of ablation study for different modules and processes in FactoScalpel.

Control Code	QAFactEval	SummaC	ClozE	DAE	FactCC	R-1	R-2	R-L
Low	20.51	13.16	75.78	**69.40**	**28.63**	40.90	17.41	32.66
Medium	20.51	13.22	75.79	69.32	28.43	**40.93**	17.45	32.17
High	**20.63**	**13.45**	**75.93**	69.35	28.13	40.92	**17.48**	**32.74**

Table 4: Results of performance for applying different control codes.

5.6 Ablation Study

We perform an ablation study via continuously removing different modules and processes in FactoScalpel. As shown the results in Table 3, each component plays a role in improving the factual consistency of the model-generated summaries. Specially, training on SummDSC-base is found to be especially important. To further study the effect of FactoScalpel module, we train a basic BART model solely on the SummDSC-base dataset. We observe a marked decline in performance without the support of FactoScalpel module. This suggests that FactoScalpel module and SummDSC-base are not just individually important but also synergistic, each enhancing the effectiveness of the other in generating faithful summaries.

Additionally, we also conduct an ablation study to assess the impact of different control codes, with the results presented in Table 4. We note a trend that the effectiveness of different control codes ranks as "High > Medium > Low". However, the overall influence is not significant. Through our analysis for summaries generated by FactoScalpel with different control codes, we find that 67.42% of the summaries are out control of control codes and same to each other. This result suggests two potential issues: the criteria for categorizing the Control Codes may be ambiguous, and the method of integrating Control Codes into the Knowledge Router might lack clarity.

5.7 Case Study

In Table 5-7 within the Appendix B, we provide several examples that illustrate the performance of various models in generating fact-aware summaries. These examples include the documents, the golden summaries, and the summaries generated by different summarization models. Our analysis reveals that both the CoFE and FactScalpel are effective, with FactScalpel exhibiting fewer instances of incorporating incorrect external hallucinations. Compared to FASum, which is also based on knowledge injection method, FactScalpel demonstrates superior adherence to the core content of the documents. This is particularly evident in Examples 1, 3, and 5, where FASum tends to stray from the main narrative. However, we also note that the current fact-aware summarization models struggle with producing coherent summaries for scenarios that involve complex logic or intertwined narratives, such as the competition outcomes between teams highlighted in Examples 3 and 4, and the challenges of disentangling multiple events as seen in Example 5.

6 Conclusion

In this paper, we propose a factual attribution algorithm and firstly verify the neurons related to facts tend to be located in higher-level FFNs, and the generation of intrinsic and extrinsic hallucinations is closely related to FFNs. Based on the attribution results, we propose a fact-aware summarization model FactoScalpel with Knowledge Bank and router-controlled mechanism. We demonstrate the effectiveness and superiority of FactoScalpel which effectively improves the factual consistency of generated summaries in an interpretable manner.

Limitations

Our study revealed that control codes had a minimal impact on our approach. Additionally, there is a need to further refine FactoScalpel, particularly in enhancing its capacity to handle intrinsic hallucinations and to improve its logical reasoning capabilities.

Ethics Statement

The models and datasets used in this paper are all open-source and do not contain any sensitive information.

Acknowledgements

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. arXiv preprint arXiv:2204.10290.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. arXiv preprint arXiv:2109.09784.

Shuyang Cao and Lu Wang. 2021a. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization. arXiv preprint arXiv:2109.09209.

Shuyang Cao and Lu Wang. 2021b. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5935–5941.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Damai Dai, Wenbin Jiang, Qingxiu Dong, Yajuan Lyu, and Zhifang Sui. 2023. Neural knowledge bank for pretrained transformers. In CCF International Conference on Natural Language Processing and Chinese Computing, pages 772–783. Springer.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022a. Calibrating factual knowledge in pretrained language models. arXiv preprint arXiv:2210.03329.

Yue Dong, John Wieting, and Pat Verga. 2022b. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. arXiv preprint arXiv:2204.13761.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. arXiv preprint arXiv:2010.08014.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3592–3603, Online. Association for Computational Linguistics.

Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. 2022. Questioning the validity of summarization datasets and improving their factual consistency. arXiv preprint arXiv:2210.17378.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S Weld, Iz Beltagy, and Doug Downey. 2022a. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. arXiv preprint arXiv:2203.08436.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022b. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 555–571, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics, 10:163–177.

Hwanhee Lee, Cheoneum Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Juae Kim, and Kyomin Jung. 2022. Factual error correction for abstractive summaries using entity retrieval. arXiv preprint arXiv:2204.08263.

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

Yiyang Li, Lei Li, Dingxin Hu, Xueyi Hao, Marina Litvak, Natalia Vanetik, and Yanquan Zhou. 2024. Improving factual error correction for abstractive summarization via data distillation and conditional-generation cloze. arXiv preprint arXiv:2402.08581.

Yiyang Li, Lei Li, Marina Litvak, Natalia Vanetik, Dingxin Hu, Yuze Li, and Yanquan Zhou. 2023. Just cloze! a novel framework for evaluating the factual consistency faster in abstractive summarization.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. arXiv preprint arXiv:2105.04623.

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

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? arXiv preprint arXiv:1909.01066.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In International Conference on Machine Learning, pages 4596–4604. PMLR.

Arvind Krishna Sridhar and Erik Visser. 2022. Improved beam search for hallucination mitigation in abstractive summarization. arXiv preprint arXiv:2212.02712.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. arXiv preprint arXiv:2112.08713.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. arXiv preprint arXiv:2210.13210.

Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. 2022a. Improving faithfulness by augmenting negative summaries from fake documents. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11913–11921.

Tianshu Wang, Faisal Ladhak, Esin Durmus, and He He. 2022b. Improving faithfulness by augmenting negative summaries from fake documents. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11913–11921, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhao Wu, Wei Li, Jiachen Liu, Xinyan Xiao, Ziqiang Cao, Sujian Li, and Hua Wu. 2022. Frsum: Towards faithful abstractive summarization via enhancing factual robustness. arXiv preprint arXiv:2211.00294.

Wang Xu and Tiejun Zhao. 2022. Jointly learning guidance induction and faithful summary generation via conditional variational autoencoders. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 2340–2350.

Yunzhi Yao, Shaohan Huang, Li Dong, Furu Wei, Huajun Chen, and Ningyu Zhang. 2022. Kformer: Knowledge injection in transformer feed-forward layers. In CCF International Conference on Natural Language Processing and Chinese Computing, pages 131–143. Springer.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022. Improving the faithfulness of abstractive summarization via entity coverage control. arXiv preprint arXiv:2207.02263.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Enhancing factual consistency of abstractive summarization. arXiv preprint arXiv:2003.08612.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 718–733, Online. Association for Computational Linguistics.

A Details of Factual Attribution

A.1 Factual Attribution Scores

For factual attribution, we define the document $\boldsymbol{X} = [x_1, x_2, \ldots, x_N]$, summary

$\boldsymbol{Y} = [y_1, y_2, \ldots, y_M]$ and factual factors $\boldsymbol{f} = [f_1, f_2, \ldots, f_K]$, each of f_i can be represented by a paragraph of text in the summary $\boldsymbol{Y}$, that is, $f_j = \left[y_{t_j}, y_{t_j+1}, \ldots, y_{t_j+Q}\right]$, where t_j represents the actual coordinate of the factual factor f_j and Q represents the length of the factual factor. From this, we can construct the conditional probability of factual factor generation as follows:

$$P\left(f_j \,|\, \boldsymbol{X}, \boldsymbol{Y}_{<t_j}, \alpha\right) = \prod_{t=t_j}^{t_j+Q} P\left(h_t \,|\, \boldsymbol{X}, \boldsymbol{Y}_{<t_j}, \alpha\right) \tag{10}$$

where α is used to control the activation value of neurons.

Since the attribution of a single word has been changed to attribution of multiple words at this time, in addition to the two dimensions of layer number and neuron number, a time step dimension will be expanded. Therefore, for the i-th neuron in the l-th layer at the $(t_j + q)$-th time step, its attribution score will be given by the following formula:

$$\mathrm{Attr}\left(w_{t_j+q,i}^{(l)}\right) = \bar{w}_{t_j+q,i}^{(l)} \int_0^1 \frac{\partial F\left(\alpha; \bar{w}_{t_j+q,i}^{(l)}\right)}{\partial w_{t_j+q,i}^{(l)}} d\alpha \tag{11}$$

where $F\left(\alpha; \bar{w}_{t_j+q,i}^{(l)}\right) = P\left(f_j \,|\, \boldsymbol{X}, \boldsymbol{Y}_{<t_j}, \alpha\right)$. In order to simplify the computation, we further introduce the Riemann approximation method to convert continuous integral into the summation of N_{bin} discrete values, as follows:

$$\mathrm{Attr}\left(w_{t_j+q,i}^{(l)}\right) = \frac{\bar{w}_{t_j+q,i}^{(l)}}{N_{bin}} \sum_{k=1}^{N_{\text{bin}}} \frac{\partial F\left(\frac{k}{N_{bin}}; \bar{w}_{t_j+q,i}^{(l)}\right)}{\partial w_{t_j+q,i}^{(l)}} \tag{12}$$

We get the factual attribution scores.

A.2 Identify Knowledge Neurons

We extract the target knowledge neurons based on the factual attribution score. We specify a parameter α as filtering threshold, and only when the neuron attribution score is higher than the threshold, we consider it as relevant to the word y_{t_j+q}. The l-th layer neuron set of the $(t_j + q)$-th time step is as follows:

$$w_{t_j+q}^{(l)} = \Big\{ w_{t_j+q,i}^{(l)} \mid \mathrm{Attr}\left(w_{t_j+q,i}^{(l)}\right) \geq \alpha \cdot \max(\mathrm{Attr}),\ 0 < \alpha < 1 \Big\} \tag{13}$$

To further find neurons related to factual factors, we give the hypothesis that words in the same factual factor share the same neurons. We set a threshold β to limit the coverage of neurons at each time step, that is, only when the proportion of the time step length of the neuron to the total time step length of the factual factor is greater than β, it is considered a neuron related to the factual factor f_j. The set of l-th layer knowledge neurons filtered out is as follows:

$$w^{(l)} = \left\{ w_i^{(l)} \,\middle|\, \frac{\sum_{t=t_j}^{t_j+Q} \mathbb{I}\left(w_{t,i}^{(l)} \in w_t^{(l)}\right)}{Q} \geq \beta\ ,\ 0 < \beta < 1 \right\} \tag{14}$$

$\mathbb{I}(*)$ is a binary function. If the condition is met, the value is 1, otherwise it is 0.

B Examples

Document #1	(...)One group was asked to begin an exercise regimen of walking around a track for 40 minutes a day, three days a week, while the others were limited to doing simple stretching and toning exercises. (. . .) Professor Erickson said: "We think of the atrophy of the hippocampus in later life as almost inevitable. But we've shown that even moderate exercise for one year can increase the size of that structure. (...)
Golden Summary	Walking for 40 minutes a few times a week is enough to preserve memory and keep ageing brains on top form, research shows.
BART-large	Regular exercise, even for 40 minutes a day, may protect the brain against decline as the ageing process ages, a study suggests.
FASum	People in the US are more likely to have dementia than previously thought, according to a new study.
CLIFF	Regular exercise in middle age may help protect the brain from dementia, a study suggests.
PINOCCHIO	Regular exercise in the 60s may protect the brain against decline as it ages, a study suggests.
CoFE	Even moderate exercise can protect the brain from memory loss in later life, a study suggests.
FactoScalpel	Regular exercise may help protect the brain from the decline in memory seen in many forms of ageing, a study suggests.
Document #2	(...)"What time does your website close?" in Poole and "Can I exercise my kestrel on your tip?" in Nottingham are among the top ten bizarre calls of the last year. (. . .) These light-hearted examples of calls highlight the variety of issues that councils deal with every day (...)
Golden Summary	Questions on everything from James Bond villains to Great British Bake Off cake tins have been received by council call centres in England.
BART-large	"How do I pay my bills?"
FASum	Councillors in Wales have been speaking to the Local councils over the past year.
CLIFF	People have been sending in so many bizarre phone calls that a mouse race, a mouse hunt and a kestrelentless seagull are among the most bizarre calls to councils in England and Wales, the Local Government Association says.
PINOCCHIO	A mouse race, a kestrel and the time of day trips to the dentist were among the "slightly amusing" queries received by councils in England and Wales last year, says the Local Government Association.
CoFE	Callers to councils in England and Wales have been given the chance to test their knowledge of some of the most common questions they receive.
FactoScalpel	"What time does your website close?" and "Can I exercise my kestrel on your tip?" have been among the "bizarre" council calls received in the past year.

Table 5: Examples for generated summaries. Where green words indicate correct factual factors, red ones indicate incorrect factual factors, orange ones indicate the external illusions that appear in the golden summary and underline ones indicate summaries with poor summarization ability.

Document # 3	(...)Although there is one fixture remaining Muckamore clinched Section A of the Lagan Valley Steels Twenty/20 Cup following a four wicket win over Lisburn. (...) On Saturday, Challenge Cup holders CSNI were well beaten by CIYMS who join Instonians, Lisburn and Muckamore in the semi-finals. (...) The other semi-finalists are Fox Lodge, Bready and Ballyspallen. (...)
Golden Summary	It was a disastrous afternoon on Sunday for North-West teams as all three lost in the second round of Irish Senior Cup so the quarter-finals will be contested by four from the NCU and four from Leinster.
BART-large	Muckamore clinched the Section A title in the Inter-Provincial Twenty/20 Cup after a nine-wicket win over Coleraine while CSNI secured the Section B title with a five-wickets win over Instonians.
FASum	Warwickshire's county champions Cup hopes were dealt a blow as they were bowled out of the quarter-finals of the intercontinental Cup at the rds.
CLIFF	Pembroke Wanderers will face the Hills and with the winners of the Northern Ireland Inter-Provincial T20 Cup semi-finals on Friday.
PINOCCHIO	Muckamore clinched Section A of the Lagan Valley Steels Twenty/20 Cup and CSNI progressed to the semi-finals after wins over Coleraine and Pembroke.
CoFE	Pembroke and Bready will meet in the semi-finals of the Lagan Valley Steels Twenty/20 Cup on Sunday.
FactoScalpel	Muckamore and CSNI have qualified for the semi-finals of the Lagan Valley Steels Twenty/20 Cup.
Document # 4	(...)Alex Schalk's header opened the scoring after eight minutes and Brian Graham's last-minute lob clinched victory. But Kilmarnock were unable to change their fate in between as they slipped to an eighth home Premiership defeat. They are eight points clear of Dundee United, who host Aberdeen on Wednesday, but have played two games more. (...)
Golden Summary	A hard-fought victory sent Ross County climbing to fourth place in the Premiership and increased the sense of alarm at Rugby Park.
BART-large	Ross County moved to within a point of second-placed Aberdeen with victory over Kilmarnock at Rugby Park.
FASum	Kilmarnock moved out of the Premiership relegation zone with a hard-fought win over Ross County at mcdiarmid Park.
CLIFF	Ross County moved to within a point of second-placed Aberdeen with a hard-fought win over Kilmarnock at Rugby Park.
PINOCCHIO	Ross County closed the gap at the top of the Premiership to eight points as they inflicted Kilmarnock's first home defeat of the season.
CoFE	Ross County kept their Premiership title hopes alive with a hard-fought victory over Kilmarnock at Rugby Park.
FactoScalpel	Kilmarnock's Premiership relegation worries continued as they were held to a goalless draw by Ross County at Rugby Park.

Table 6: Continuation of Table 5.

Document # 5	(...)Gary Neville, from Cupar Way, Belfast, raised Â£930 at a charity fundraising music night he organised last year. He was prosecuted for failing to give the cash to the Neuroblastoma Children's Cancer charity. (...) The Knox family set up the Oscar Knox Appeal and still use the networking sites to raise awareness and funds for organisations helping children with Neuroblastoma. (...)
Golden Summary	A disc jockey has been jailed for eight months for stealing money from a children's cancer charity linked to the Oscar Knox appeal.
BART-large	A man who used the death of a child's cancer battle to raise money for a charity has been jailed for eight months.
FASum	The family of a County Antrim boy who died of cancer after taking part in a charity fundraising campaign have apologised for his death.
CLIFF	A man who used money raised at a cancer charity to pay for the cancer treatment of a young boy has been jailed for eight months.
PINOCCHIO	A man who failed to give cancer money raised at a concert in memory of Oscar Knox has been jailed for eight months.
CoFE	A man who used Facebook to raise money for a cancer charity has been jailed for eight months at Londonderry Magistrates Court.
FactoScalpel	A man who used a charity fundraising event to raise money for a child's cancer charity has been jailed for eight months.

Table 7: Continuation of Table 6.