A Lightweight Constrained Generation Alternative for Query-focused Summarization

Zhichao Xu
zhichao.xu@utah.edu
University of Utah
Salt Lake City, Utah, United States

Daniel Cohen
daniel.cohen@dataminr.com
Dataminr, Inc.
NYC, NY, United States

ABSTRACT

Query-focused summarization (QFS) aims to provide a summary of a document that satisfies information need of a given query and is useful in various IR applications, such as abstractive snippet generation. Current QFS approaches typically involve injecting additional information, e.g. query-answer relevance or fine-grained token-level interaction between a query and document, into a finetuned large language model. However, these approaches often require extra parameters & training , and generalize poorly to new dataset distributions. To mitigate this, we propose leveraging a recently developed constrained generation model Neurological Decoding (NLD) as an alternative to current QFS regimes which rely on additional sub-architectures and training. We first construct lexical constraints by identifying important tokens from the document using a lightweight gradient attribution model, then subsequently force the generated summary to satisfy these constraints by directly manipulating the final vocabulary likelihood. This lightweight approach requires no additional parameters or finetuning as it utilizes both an off-the-shelf neural retrieval model to construct the constraints and a standard generative language model to produce the QFS. We demonstrate the efficacy of this approach on two public QFS collections achieving near parity with the state-of-the-art model with substantially reduced complexity.

CCS CONCEPTS

• **Information systems**; • **Computing methodologies** → **Natural language generation**;

KEYWORDS

Query-focused Summarization, Constrained Generation

ACM Reference Format:
Zhichao Xu and Daniel Cohen. 2023. A Lightweight Constrained Generation Alternative for Query-focused Summarization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591936

1 INTRODUCTION

In modern search systems, users are often presented with short *snippets* of a candidate document on their search results page. This snippet serves as a critical element in helping users determine whether a document satisfies their information needs without requiring them to invest additional time. The effectiveness of a snippet largely depends on its ability to accurately and concisely capture the relevant information from the corresponding document in just a few lines of text [5, 23].

This task of query-focused summarization (QFS) snippet generation, commonly referred to as query-biased summarization [31] or abstractive snippet generation [4], aims to construct a summary that succinctly addresses the information need of a query by extracting essential information from a document. Traditionally, QFS has used extractive methods that rely on the most relevant spans of text from a candidate document based on the prevalence of query terms [2, 32]. Although efficient, this extractive approach is constrained by the format of the original document, with the effectiveness of the snippet heavily dependent on the length and information density of the candidate document [31]. Moreover, this paradigm limits the possibility of personalization or multiple-document QFS.

With the advent of large language models (LM), a new paradigm has emerged that attempts to address the limitations of extractive snippet generation. These LM-based approaches directly generate abstractive snippets that do not necessarily appear anywhere in the original document [4, 14, 23]. While these methods hold promise, successful application is nontrivial. They often require specific architectures with additional parameters to incorporate the relevance signal, along with extensive fine-tuning. Moreover, a significant challenge with natural language generation, including QFS, is the problem of hallucination, where the model confidently generates false information [11, 19]. This can lead to unreliable snippets that do not accurately reflect the content of the original document.

In this paper, we propose a novel lightweight alternative approach for QFS, relevance-constrained QFS, that achieves near parity with current state-of-the-art methods with significantly reduced complexity. Our approach does not require training an additional model or adding more parameters to the existing one. Instead, we demonstrate that QFS can be effectively and efficiently achieved by constraining a general language model (LM) to summarize the document with predefined lexical constraints. We hypothesize that QFS can be effectively achieved by enforcing the language model to summarize with predefined lexical constraints – i.e, constraining a pretrained LM to favor the most important terms for the relevance of the document. While the notion of constrained generation [18] has emerged as a way to combat hallucination in other natural language generation tasks such as commonsense generation, constrained machine translation, conversation [35] and table to text generation [17], we show the unique advantages it possesses in the case of abstractive QFS.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591936

To achieve this, we first identify the most critical tokens from the ranking model using the gradient signal of each token [26] as these salient terms capture the most important aspects of the document's relevance to the query. We then convert these tokens into predicate logic constraints and use them as input to a version of constrained generation, Neurological Decoding [18]. By constraining the LM to simultaneously satisfy these constraints and maintain fluency, we generate an abstractive summary that is optimized for relevance to the query. This approach allows us to effectively generate snippets without requiring additional complex modules or training methods, making it a lightweight yet effective alternative to the current state-of-the-art method.

Our experiments on two benchmark snippet generation datasets [12, 20] demonstrate that this application of relevance-constrained QFS achieves comparable results to the current state-of-the-art method, suggesting a promising alternative perspective to the snippet generation task.

2 RELATED WORK

Query-focused Summarization: To generate a query-focused summary, several studies used an additional query-attention mechanism. QR-BERTSUM-TL [13] incorporates query relevance scores into a pre-trained summarization model. Su et al. [29] propose merging the representation of an answer span predicted by a separate QA model into the Seq2Seq model's training and inference process to enforce the summary's coherence w.r.t. the query. QSG Transformer [23] suggests using a separate graph neural network model to learn per-token representations and fuse them to the Seq2Seq model to effectively generate a QFS. These mechanisms can be viewed as enforcing soft semantic constraints during the generation process, and requires additional modules and parameters to function effectively. We opt for a different approach, i.e. explicitly enforcing lexical constraints during the generation process, without the additional machinery that is necessary to handle the soft semantic constrains.

Constrained Generation (or Conditional Generation) is a family of natural language generation (NLG) methods that aim to generate natural language including/excluding a set of specific words, i.e. lexical constraints. The NLG domain recipe leverages pre-trained large language models (LLM) finetuned on specific datasets [7]. However, as pointed out by Lu et al. [18], such models only finetuned in an end-to-end manner do not learn to follow the underlying constraints reliably even when supervised with large amounts of training examples. Therefore, a line of works [1, 10, 17, 18] in constrained generation proposes to explicitly modify the likelihood of next word prediction in the generation stage, such that the predefined lexical constraints can be better satisfied.

3 RELEVANCE-CONSTRAINED QFS

Problem Formulation: Given a query-document pair (q, d), our task is to generate an abstract summarization s, which addresses the information need of the query.

We propose addressing this problem by leveraging a relevance-constrained generation. In this section, we first introduce how we construct the set of constraints used by the language model to generate the abstract summary. We then present the constrained generation process itself.

Identifying Constraints: In order to identify the most effective constraints for QFS, we first assume that each candidate document is relevant to the query. We then use a pointwise cross-entropy loss, $\mathcal{L}$, to identify how each token contributes to the relevance of the document. To achieve this, we use a saliency based mapping approach to quantify this impact as gradient-based attribution methods have been widely adopted in existing NLP literature [8, 25, 34].

Formally, denote an input sequence $(w_1, w_2, \cdots, w_n)$, where w_i is the i-th token; and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ is a sequence of corresponding static token embeddings. Let $f(\cdot)$ be a function that takes $\mathbf{x}$ as input and outputs a prediction logit, e.g., a transformer-style model with classification head. The gradients w.r.t. each input token w_i can be regarded as each token's contribution, or *saliency*, to the final prediction $f(\mathbf{x})$. We denote this per token gradient vector as $\mathbf{a} = (a_1, a_2, \cdots, a_n)$, which is the normalized saliency across all tokens,

$$a_i = \frac{g(\nabla_{\mathbf{x}_i}\mathcal{L},\ \mathbf{x}_i)}{\sum_{j=1}^{n} g(\nabla_{\mathbf{x}_j}\mathcal{L},\ \mathbf{x}_j)} \tag{1}$$

where $\mathcal{L}$ denotes the loss between $f(\mathbf{x})$ and label $y = 1$, and $g(\cdot, \cdot)$ is the saliency function.

While there exists various methods to estimate the saliency via $g(\cdot, \cdot)$ [8, 27, 28, 30, 34], we adopt `InteGrad` [30], as it is robust to input perturbations [34]. Specifically, `InteGrad` sums the gradients along the path from a baseline input $\mathbf{x}'_i = \mathbf{0}$ to the actual input $\mathbf{x}_i$:

$$g(\nabla_{\mathbf{x}_i}\mathcal{L}, \mathbf{x}_i) = (\mathbf{x}_i - \mathbf{x}'_i) \times \sum_{k=1}^{m} \frac{\partial f(\mathbf{x}'_i + \frac{k}{m} \times (\mathbf{x}_i - \mathbf{x}'_i))}{\partial\, \mathbf{x}_i} \tag{2}$$

where m is the number of steps to interpolate the input $\mathbf{x}_i$ and $\times$ denotes dot product; thus $g(\nabla_{\mathbf{x}_i}\mathcal{L}, \mathbf{x}_i)$ is a scalar indicating saliency of token w_i before normalization (Eq. 1). In our implementation, we follow the original setup in [30] and set m to 10 steps.

We note that any differentiable retrieval function can be used in place of $f(\cdot)$ within this framework. In this paper, we use a standard DistilBERT document reranker trained on MS MARCO using a cross-entropy loss [3, 9, 21, 22].

In our preliminary experiments, we observed that the saliency scores are often noisy, attributing gradients to stopwords and/or punctuations. Therefore, we filter out the stopwords and punctuations in a post hoc manner and only keep the top-3 important tokens from document d to construct the actual decoding constraints C.

Constructing Constraints: Having identified the most salient tokens, we construct the lexical constraints in a format appropriate for constrained generation, Conjunctive Normal Form,

$$C = \underbrace{(D_1 \vee D_2 \vee \cdots \vee D_i)}_{C_1} \wedge \cdots \wedge \underbrace{(D_k \vee D_{k+1} \vee \cdots \vee D_n)}_{C_m}$$

where each single D_i denotes one single positive or negative constraint, which we refer to as a *literal*; and the logical disjunction of literals is referred to as a *clause*, e.g. C_1 to C_m. In our implementation, we construct 3 clauses with each clause initially consisting of a single literal. We then expand each clause by all possible forms of the original token via WordForms[1]. An example of this logic

[1] https://github.com/gutfeeling/word_forms

corresponding to Row 1, Table 2 is represented as

$$C = \underbrace{(\text{private} \vee \ldots \vee \text{privatization})}_{C_1} \wedge \underbrace{(\text{health} \vee \ldots \vee \text{healthy})}_{C_2} \wedge \underbrace{(\text{standard} \vee \ldots \vee \text{standards})}_{C_3}$$

Constrained Generation: At inference time, we run a simplified version of the Neurological Decoding (NLD) algorithm using the set of constraints C acquired from Section 3. As we do not use negative constraints in QFS, i.e. we do not avoid certain tokens, we consider only two states within the original NLD algorithm: *reversible unsatisfaction* where an unsatisfied logical clause with a positive literal can be satisfied at a future point and *irreversible satisfaction* where a positive literal will remain satisfied. This predicate logic is then applied within a conventional beam search during generation.

At timestep t, the simplified algorithm performs three individual steps when filling in beam candidates: *Pruning*, *Grouping*, and *Selecting*. Pruning filters out candidates that are of low likelihood or satisfy fewer clauses; Grouping implicitly constructs the power set of all irreversible satisfied clauses, leading to at most $2^{|C|}$ groups; and Selecting populates the beam with candidates within each group that are most likely to satisfy remaining reversible unsatisfied clause C_j by modifying the likelihood. Specifically, within each group, the likelihood is modified by the NLD score function:

$$L = P_\theta(y_t|y_{<t}) + \lambda \max_{\mathbb{I}(C_j)=0} \frac{|\hat{D}_i|}{|D_i|} \tag{3}$$

where P_θ is the likelihood of the LM generating token y_t, $\mathbb{I}(C_j)$ indicates whether clause C_j has been satisfied or not, $\frac{|\hat{D}_i|}{|D_i|}$ is the overlap between the ongoing generation and the partially satisfied literal D_i, e.g. $\hat{D}_i$ ="apple" and D_i ="apple tree" yields 0.5, and λ = 0.1 acts as the hyperparameter. Intuitively, this score modification favors candidates moving toward fully satisfying a positive literal within an unsatisfied clause with λ controlling the strength of this signal. After this explicit likelihood modification, we visit each group and select the highest scoring candidate in rotation until the beam is filled. After this process is complete, we select the beam candidate with highest score and proceed to generating the next token at $t + 1$. Although the group construction suggests a high-complexity runtime, implicit construction results in this algorithm having the same runtime complexity as standard beam search [18].

We use BART [14] and T5 [24] for fair comparison with existing methods as the generating LM for abstractive QFS. As there exist no additional parameters or modules for this method, details of these backbone LMs are discussed in Section 4.

4 EXPERIMENTAL SETUP

Datasets: Following previous works [23, 29], we adopt Debatepedia [20] and PubMedQA [12] to benchmark the effectiveness of the proposed relevance-constrained generation method. Debatepedia dataset is collected by Nema et al. [20] from 663 debates of 53 diverse categories in an encyclopedia of debates and consists of 12K/0.7K/1.0K query-document summarization triplets (q, d, s). PubMedQA is a long-form abstractive question-answering dataset from the biomedical domain with the contexts available. We use the standard train test split from the original datasets.

Table 1: Results on test set, including ROUGE-1, ROUGE-2 and ROUGE-L, baseline results (the first section) are from [23]; *Italic* indicates the best performing system in literature. † denotes the constrained method significantly better than its unconstrained counterparts with paired t-test at 0.05 level

Model	Debatepedia			PubMedQA		
	R-1	R-2	R-L	R-1	R-2	R-L
Transformer	41.7	33.6	41.3	30.4	8.4	22.3
SD2	41.3	18.8	40.4	32.3	10.5	26.0
CSA Transformer	46.4	37.5	45.9	-	-	-
QR-BERTSUM-TL	48.0	45.2	57.1	-	-	-
MSG	-	-	-	37.2	14.8	**30.2**
BART-QFS	59.0	44.6	57.4	-	-	-
QSG BART	**64.9**	**52.3**	**63.3**	**38.4**	**17.0**	29.8
T5	22.5	7.1	19.7	38.0	15.3	28.2
Constrained-T5	32.2†	12.5†	28.2†	36.4	16.0†	28.7†
- Rel. Improv. (%)	+43.1	+76.1	+43.2	−4.3	+4.6	+1.8
BART	58.1	43.6	56.8	38.1	15.7	27.2
Constrained-BART	**62.9**†	**50.1**†	**61.5**†	**39.2**†	**17.1**†	**30.1**†
- Rel. Improv. (%)	+8.3	+14.9	+8.3	+2.9	+8.9	+10.6

Compared Methods: To evaluate the performance of the proposed relevance-constrained generation method, we introduce the following baseline methods in order of increasing complexity:

- **End-to-End approaches**: Transformer [33], BART [14] and T5 [24] are finetuned for Seq2Seq summarization. These LMs additionally act as the backbone LM for the proposed relevance-constrained QFS approach, i.e. Constrained-BART and Constrained-T5 such that the results are directly comparable. In this configuration, there are no constraints during the generation process.
- **Improved query-document cross attention**: SD2 [20] adds additional cross attention between query and document encoder, then uses the combined representation for generation. CSA Transformer [36] adds conditional self-attention layers originally designed for conditional dependency modeling to the Seq2Seq model.
- **Incorporated query-document relevance**: QR-BERTSUM-TL [13] injects query relevance scores into pretrained Seq2Seq summarization model; MSG [6] utilizes query relevance and interrelation between sentences of the document for fine-grained representation. Similarly, BART-QFS [29] also uses a pre-trained QA model to determine answer relevance in the document and injects this information into the Seq2Seq LM model.
- **Additional module utilization**: QSG-BART [23] utilizes an additional graph neural network module to model token-level interaction between query and document, and injects this information into Seq2Seq model. It reaches state-of-the-art performance on the QFS task, but requires additional parameters and training.

Evaluation Metrics: We evaluate the effectiveness of Constrained QFS with ROUGE-1, ROUGE-2, and ROUGE-L [15] for fair comparison to existing works [23, 29].

Implementation Details: We adopt an off-the-shelf Cross Encoder model[2] as our saliency model. We identify the top-3 important tokens with Eq.2 and construct constraints as $C = C_1 \wedge C_2 \wedge C_3$.

We experiment with two pre-trained Seq2Seq models as the base generator, T5 [24] and BART [14]. Different from previous works BART-QFS and QSG BART [23, 29], we do not warm start

[2]https://huggingface.co/brutusxu/distilbert-base-cross-encoder-first-p

Table 2: Sample qualitative study on Debatepedia dataset; tokens are marked salient and included in constraints set C.

Query	Document	BART Generation	Constrained-BART Generation	Golden Reference
privatization: is water a resource that should be owned by private companies versus a global commons?	**private** companies are profit-maximizing entities that often view environmental **health** and safety **standards** as obstructive to their profit interests. this is a problem particularly in the context of water which is so fundamentally important to the environment health and life.	environmental and health standards are often violated by public owners of water	environmental and **health** **standards** are often violated by water **privatization** companies.	environmental and health standards are often violated by private ownership of water
competition: does a public option increase competition?	"the case against: the public plan will unfairly crowd out private coverage". heritage foundation. july 28 2009: "it 's simply impossible to believe the claims by sen. charles schumer (d-n.y.) and others that **congress** really will do nothing to disrupt the level playing field by favoring the public plan. with congress as both umpire and a team manager one thing is clear: it will favor its own team. the result is the **public** plan will unfairly crowd out **private** coverage.	congress will not favor pri -vate insurer over private insurers	**congress** will not favor **private** insurer over **public** insurers	government will favor public insurance; no level playing -field

Table 3: Effect of the source of constraints to QFS performance on Constrained BART. † denotes significantly better than the other two methods with paired t-test at 0.05 level.

Constraints	Debatepedia			PubMedQA		
	R-1	R-2	R-L	R-1	R-2	R-L
Query-only	61.4	48.4	59.9	39.0	16.9	29.9
Document-only	**62.9**†	**50.1**†	**61.5**†	**39.2**	**17.1**	**30.1**
Query+Document	61.5	48.4	60.1	38.9	17.0	29.7

BART or T5 by pre-finetuning on existing abstractive summarization datasets; instead we only finetune them on our target datasets Debatepedia and PubMedQA. For T5, we format the input as `Summarize: Document:` d `Question:` q: and finetune the model weights on each dataset's training set with golden references. At inference time, we use the same input format and finetuned model weights for relevance-constrained generation/generation. For BART, we format the input as `[CLS]` d `[SEP]` q `[EOS]`, where `[CLS]`, `[SEP]`, `[EOS]` are special tokens indicating start, separate and end of sequence, then we finetune and generate text in a similar fashion to T5. For both models, we finetune with AdamW optimizer [16], learning rate $2e-5$ and early stop after no improvements on the dev set for three consecutive epochs. We make our code publicly available at https://github.com/zhichaoxu-shufe/Constrained-QFS.

5 RESULT AND ANALYSIS

We address two RQs in this section:

- **RQ1**: How competitive is the proposed constrained generation method in terms of performance compared to baselines?
- **RQ2**: How does constrained generation affect QFS performance?

To answer **RQ1**, shown in Table 1, we observe that the relevance-constrained methods achieve competitive performance on two datasets. On Debatepedia dataset, Constrained-BART achieves near parity with the current state-of-the-art system and substantially outperforms all other baselines. This result is particularly interesting given the reduced complexity Constrained-BART. On the PubMedQA dataset, Constrained-BART achieves slightly better performance than QSG BART. A possible explanation for this improved performance might be the length of the documents in PubMedQA, where the relevance-constrained process results in a more consistent snippet. We therefore conclude that the proposed relevance-constrained generation paradigm can achieve competitive performance without additional parameters or finetuning.

To answer **RQ2**, we specifically draw a comparison between the proposed methods and their unconstrained baselines, which were finetuned end-to-end and generated QFS without constraints. In the second section of Table 1, we observe that the proposed constrained generation methods consistently outperform their unconstrained counterparts across different datasets and backbone LMs. For instance, on the Debatepedia dataset, Constrained-BART outperforms BART 14.9% in R-2. Therefore, we conclude that by adding carefully constructed constraints into the generation stage, the performance of the QFS task can be significantly improved without modifying the backbone LMs.

Qualitative Analysis: We show two examples in Table 2. In the first example, the BART generation hallucinates "public owners" that is not faithful to the document; however, Constrained-BART is able to successfully summarize the document as C contains "privatization". In the second example, despite the underspecified query, the saliency model still extracts critical tokens, which are able to aid in the generation of a meaningful summary.

Ablation Study: In Table 3 we study the effect of different sources of constraints. Query-only denotes that the top-3 important tokens are from the query, and vice versa for Document-only and Query+Document. We observe that on the Debatepedia dataset, Document-only constraints significantly outperforms the other two approaches, while on PubMedQA this improvement is minor. After manual examination, we find that the golden references in Debatepedia dataset overlap more with documents compared to queries, while PubMedQA does not adhere to this trend.

6 CONCLUSION AND FUTURE WORK

In this work, our lightweight relevance-constrained generation approach achieves competitive performance compared to the state-of-the-art method, and it can easily generalize to new domains provided the existence of an effective retrieval model to guide the constraint construction. Our future work may involve investigating the effectiveness and summarization faithfulness/factuality of this approach in real world IR systems.

7 ACKNOWLEDGEMENT

Zhichao Xu is supported partially by NSF IIS-2205418 and NSF DMS-2134223. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576* (2016).

[2] Hannah Bast and Marjan Celikik. 2014. Efficient index-based snippet generation. *ACM Transactions on Information Systems (TOIS)* 32, 2 (2014), 1–24.

[3] Leonid Boytsov, Tianyi Lin, Fangwei Gao, Yutian Zhao, Jeffrey Huang, and Eric Nyberg. 2022. Understanding Performance of Long-Document Ranking Models through Comprehensive Evaluation and Leaderboarding. *arXiv preprint arXiv:2207.01262* (2022).

[4] Wei-Fan Chen, Shahbaz Syed, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Abstractive snippet generation. In *Proceedings of The Web Conference 2020*. 1309–1319.

[5] Hoa Trang Dang. 2006. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*. 48–55.

[6] Yang Deng, Wenxuan Zhang, and Wai Lam. 2020. Multi-hop inference for question-driven summarization. *arXiv preprint arXiv:2010.03738* (2020).

[7] Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research* 73 (2022), 1131–1207.

[8] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781* (2018).

[9] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink training of BERT rerankers in multi-stage retrieval pipeline. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 280–286.

[10] Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138* (2017).

[11] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *Comput. Surveys* (2022).

[12] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146* (2019).

[13] Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Huang. 2020. Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models. In *Canadian conference on artificial intelligence*. Springer, 342–348.

[14] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[15] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[16] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[17] Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, et al. 2021. Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. *arXiv preprint arXiv:2112.08726* (2021).

[18] Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Neurologic decoding:(un) supervised neural text generation with predicate logic constraints. *arXiv preprint arXiv:2010.12884* (2020).

[19] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[20] Preksha Nema, Mitesh Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. *arXiv preprint arXiv:1704.08300* (2017).

[21] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.

[22] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713* (2020).

[23] Choongwon Park and Youngjoong Ko. 2022. QSG Transformer: Transformer with Query-Attentive Semantic Graph for Query-Focused Summarization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2589–2594.

[24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[25] Alexis Ross, Ana Marasović, and Matthew E Peters. 2020. Explaining nlp models via minimal contrastive editing (mice). *arXiv preprint arXiv:2012.13985* (2020).

[26] Ivan Sanchez, Tim Rocktaschel, Sebastian Riedel, and Sameer Singh. 2015. Towards extracting faithful and descriptive representations of latent variable models. *AAAI Spring Syposium on Knowledge Representation and Reasoning (KRR): Integrating Symbolic and Neural Approaches* 1 (2015), 4–1.

[27] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[28] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[29] Dan Su, Tiezheng Yu, and Pascale Fung. 2021. Improve query focused abstractive summarization by incorporating answer relevance. *arXiv preprint arXiv:2105.12969* (2021).

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[31] Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 2–10.

[32] Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E Williams. 2007. Fast generation of result snippets in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 127–134.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[34] Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of NLP models is manipulable. *arXiv preprint arXiv:2010.05419* (2020).

[35] Zhenduo Wang, Zhichao Xu, Qingyao Ai, and Vivek Srikumar. 2023. An In-depth Investigation of User Response Simulation for Conversational Search. *arXiv preprint arXiv:2304.07944* (2023).

[36] Yujia Xie, Tianyi Zhou, Yi Mao, and Weizhu Chen. 2020. Conditional self-attention for query-based summarization. *arXiv preprint arXiv:2002.07338* (2020).