Constrained Abstractive Summarization: Preserving Factual Consistency with Constrained Generation

Yuning Mao[1], **Xiang Ren**[2], **Heng Ji**[1], **Jiawei Han**[1]
[1]University of Illinois, Urbana-Champaign [2]University of Southern California
[1]{yuningm2, hengji, hanj}@illinois.edu [2]xiangren@usc.edu

Abstract

Summaries generated by abstractive summarization are supposed to only contain statements entailed by the source documents. However, state-of-the-art abstractive methods are still prone to hallucinate content inconsistent with the source documents. In this paper, we propose constrained abstractive summarization (CAS), a general setup that preserves the factual consistency of abstractive summarization by specifying tokens as constraints that must be present in the summary. We explore the feasibility of using lexically constrained decoding, a technique applicable to any abstractive method with beam search decoding, to fulfill CAS and conduct experiments in two scenarios: (1) Standard summarization without human involvement, where keyphrase extraction is used to extract constraints from source documents; (2) Interactive summarization with human feedback, which is simulated by taking missing tokens in the reference summaries as constraints. Automatic and human evaluations on two benchmark datasets demonstrate that CAS improves the quality of abstractive summaries, especially on factual consistency. In particular, we observe up to 11.2 ROUGE-2 gains when several ground-truth tokens are used as constraints in the interactive summarization scenario.

Reference: sir tom jones is to return as one of the judges on talent show **the voice uk** when it moves to **itv** next year. **Unconstrained**: pop star sir tom jones is to return to the **bbc** 's voice uk after a two-year absence. **Constrained**: singer sir tom jones is to return to **itv** 's **the voice uk** next year after a two-year absence.
Reference: syrian refugees facing their first **christmas** in wales are sure to get a " warm welsh welcome ", the first minister has said. **Unconstrained**: the first minister has said wales is **" more important than ever " in the new year**. **Constrained**: the first welsh councils to welcome refugees in the uk this **christmas** have been praised by the first minister.
Reference: a four-month consultation which could help decide the location of the uk 's first spaceport ends on monday with one site in **gwynedd** being considered. **Unconstrained**: plans for a uk spaceport and spaceport in snowdonia have been backed by **a council**. **Constrained**: plans to build a uk spaceport in snowdonia have been backed by **gwynedd** council.

Table 1: **Comparison of unconstrained and constrained summaries**. Constrained (replaced) tokens are in **green** (**red**).

1 Introduction

Abstractive summarization aims to generate condensed summaries covering salient and factual information in the source documents. Although abstractive summarization has achieved significant progress with advances in Seq2Seq learning (See et al., 2017; Paulus et al., 2017) and pre-trained language models (Liu and Lapata, 2019; Lewis et al., 2019), the generation process is unconstrained – abstractive models learn to generate summaries in a completely data-driven manner using document-summary pairs. As a consequence, they are prone to hallucinate content not entailed by the source documents (*e.g.*, producing unseen entities or unfaithful facts) (Kryscinski et al., 2019; Maynez et al., 2020). Such factual inconsistencies in the summary hinder the feasibility of abstractive methods being deployed in real-world applications.

In this paper, we propose constrained abstractive summarization (CAS), where a set of tokens are used as constraints and required to be present in the generated summary. CAS improves the factual consistency of abstractive summarization in two ways. First, we observe that the added constraints can often replace their unfaithful counterparts in the unconstrained summary of the same model and thus reduce model hallucination. For example, in Table 1, the model corrects "BBC" to "ITV" as "The Voice UK" is acquired by "ITV". Second, when adding important entities in the source document as constraints, the model is more likely to generate summaries that are focused on these factual entities (*e.g.*, “Christmas”) and more specific (*e.g.*, “a council” changed to “Gwynedd council”). To ensure the presence of the constraints in the summary, we use lexically constrained decoding (Post and Vilar, 2018), because it only functions during inference as a plug-and-play module and can be integrated into any abstractive model with beam search decoding. In this way, one can easily conduct CAS on (almost) any fine-tuned abstractive models without (often expensive) re-training.

We consider two scenarios for CAS: (1) a standard summarization scenario without human involvement, where we conduct keyphrase extraction to extract keyphrases from the source documents as constraints; (2) a human-in-the-loop interactive summarization scenario, where the human feedback is mimicked by the reference summaries and missing tokens in the reference summaries are used as constraints to guide CAS towards human preferences. We verify the effectiveness of CAS on the CNN/Daily Mail (Nallapati et al., 2016) and XSum (Narayan et al., 2018) datasets with BERTSum (Liu and Lapata, 2019) as the base abstractive model. Both automatic and human evaluations demonstrate that CAS generates summaries of higher quality, leading to improvements of both lexical overlap and factual consistency. Moreover, we show that CAS with BERTSum achieves better performance than state-of-the-art methods (Lewis et al., 2019; Zhang et al., 2020a) by simply using *one* phrase in the reference summaries as “hint” (constraint) during inference, which shows its significant benefits in interactive summarization and potentials of future development.

2 Related Work

Factual Consistency of Summarization. Recent studies (Kryscinski et al., 2019; Maynez et al., 2020) show that up to 30% summaries of state-of-the-art abstractive methods contain factual inconsistencies on CNN/Daily Mail and more than 70% on the more abstractive XSum dataset. As ROUGE (Lin, 2004) does not correlate well with factual consistency (Falke et al., 2019), model-based metrics learned by weak supervision (Goodrich et al., 2019; Kryscinski et al., 2019) are proposed to measure factual consistency explicitly. However, the performance of these metrics is still unsatisfactory and to date, there is no commonly accepted metric for the evaluation of factual consistency beyond human evaluation. Most of previous studies (Matsumaru et al., 2020; Zhu et al., 2020) improve factual consistency at the expense of lexical overlap (significantly lower ROUGE) while CAS improves both metrics simultaneously.

Constrained Generation. Constrained Generation is useful in various applications such as machine translation (Hokamp and Liu, 2017), data augmentation (Hu et al., 2019a), and review generation (Zhang et al., 2020b). Although approaches like copy mechanism (See et al., 2017) that encourage models to copy words from source documents have been widely adopted, they are often insufficient to reduce summary hallucination (Maynez et al., 2020). To our knowledge, constrained generation like CAS that requires certain tokens must be present in a summary is unexplored.

3 Constrained Abstractive Summarization

3.1 Task Formulation

We define a constraint set $\mathcal{C}$ as a set of words (phrases) $\{p_1, p_2, ...p_N\}$ where p_i is a text span of arbitrary length. Given document-reference pair (d, r) and abstractive model $\mathcal{M}$, the goal of constrained abstractive summarization (CAS) is to generate a summary s for document d using $\mathcal{M}$ *with the presence of all the text spans* in $\mathcal{C}$. We assume that an unconstrained summary s' is generated by the same $\mathcal{M}$ (with $\mathcal{C} = \emptyset$) to compare with s. An ideal constraint set $\mathcal{C}$ should have a high overlap with r to ensure its quality and low overlap with s' to bring additional information.

3.2 Constrained Generation

We use one lexically constrained decoding method, dynamic beam allocation (DBA) (Post and Vilar, 2018), for constrained generation in CAS. At a high level, DBA divides the beam during beam search to store hypotheses satisfying different numbers of constraints and adds unmet constraints at each decoding step. DBA ensures the presence of constraints by allowing [EOS] token only when all the constraints are met. We choose DBA due to its fast decoding speed over other counterparts (Hokamp and Liu, 2017) – complexity of $\mathcal{O}(1)$ in the number of constraints. In addition, DBA completely works at the inference stage and thus can be easily incorporated into different models for the evaluation of CAS, which is preferable over methods that involve training (Zhang et al., 2020b) since we want to keep the fine-tuned, often expensive abstractive models intact rather than re-train them (with a possibly different objective).

CAS is useful for improving abstractive summarization, especially on factual consistency in two scenarios. First, one can obtain the constraints automatically without human involvement. The source document is a natural choice for constraint discovery since the goal is to generate a summary consistent with it. Specifically, we extract keyphrases in the source documents and use the extracted keyphrases as the constraints (Sec. 3.3). Second, for human-in-the-loop interactive summarization (Gao et al., 2018; Shapira et al., 2020), when a model generates a summary that contains factual errors or lacks certain information, one can add the corrected or missing facts as constraints and conduct constrained generation. We simulate human preferences by adding missing tokens in the reference summary as constraints (Sec. 3.4).

3.3 CAS with Automatic (Noisy) Constraints

To obtain the constraints automatically, we perform keyphrase extraction on the source documents and conduct constraint filtering to reduce noise.

Keyphrase Extraction (KPE). The quality of constraints (keyphrases) is critical as they will be "inserted" to the summary. In preliminary experiments, we find commonly used unsupervised KPE methods (Campos et al., 2020) insufficient to provide high-quality constraints. Therefore, we use the state-of-the-art supervised KPE method, BERT-KPE (Sun et al., 2020), to extract keyphrases from the source documents. To train BERT-KPE, we take the same data split as the corresponding summarization dataset. We hypothesize that entities and noun phrases in the source document carry the most critical information regarding factual consistency and thus take them as constraints. We use spaCy (Honnibal and Montani, 2017) to find named entities and noun phrases in the reference summaries of the training set, and treat those appearing in the source documents as positive samples. All N-grams up to $N = 5$ are considered as constraint candidates for extraction.

Constraint Filtering. Not all extracted keyphrases are important (*i.e.*, appear in the reference summary) and many high-quality keyphrases have already been covered by the unconstrained summary s'. Therefore, we filter out the keyphrases appearing in s' such that only constraints otherwise not presented in the summary (*i.e.*, bringing additional information) are kept. To further reduce noise, we only take top 3 keyphrases with scores greater than 1.6 (defined by BERT-KPE) as constraints. As a result, the quality of the constraints and consequently the constrained summaries is ensured.

3.4 CAS with Manual (Ground-truth) Constraints

Interactive summarization (Gao et al., 2018; Shapira et al., 2020) that involves human interactions is drawing more attention as different readers may have different preferences or information needs even for the same document. However, human feedback for summarization is often expensive to obtain (Stiennon et al., 2020). Therefore, we compare the unconstrained summary s' with the reference r to simulate human preferences. Specifically, we assume that humans would prefer r over s' and add several text spans presented in r but not in s' as the constraints. CAS with ground-truth constraints are not only useful for interactive summarization but also help us understand the potential (upper bound) of CAS when high-quality constraints are provided. Similar pick-revise interactive cycles have proved effective in machine translation for post-editing (Cheng et al., 2016).

4 Experiments

4.1 Experimental Setup

We conduct experiments on the CNN/Daily Mail (Nallapati et al., 2016) and XSum (Narayan et al., 2018) datasets. We use one state-of-the-art abstractive model, BERTSum (Liu and Lapata, 2019), as our base model $\mathcal{M}$. For beam search, we set beam size to 10 when noisy constraints are used and 5 for ground-truth constraints unless otherwise specified.

4.2 CAS with Noisy Constraints

Automatic Evaluation. As a sanity check for keyphrase extraction, we achieve 0.756 Precision@1 and 0.489 F1@5 on CNN/Daily Mail, and 0.665 Precision@1 and 0.400 F1@5 on XSum. In Table 2, we show the comparison of the base model BERTSum and CAS. Our results are slightly different from Liu and Lapata (2019) despite using its official code and model weights. CAS outperforms BERTSum by 0.4 to 0.5 in ROUGE on CNN/Daily Mail and 0.2 to 0.3 on XSum *with statistical significance*. The greater improvement on CNN/Daily

	Method	R-1	R-2	R-L
CNN/DM	BERTSum	42.00	19.44	38.98
	CAS	**42.48**	**19.56**	**39.43**
XSum	BERTSum	38.91	16.54	31.30
	CAS	**39.19**	**16.75**	**31.56**

Table 2: **Comparison between unconstrained and constrained summarization**. Improvements are statistically significant ($p < 0.05$) under approximate randomization test and paired bootstrap resampling test.

Mail is likely because its reference summaries are more extractive in nature (Mao et al., 2020) and its extracted keyphrases are of higher quality.

Human Evaluation. We note that CAS is only applied to a small subset of samples (*e.g.*, 9.5% on XSum) due to constraint filtering and we show in Sec. 4.3 that CAS achieves significantly higher ROUGE in interactive summarization. Furthermore, previous studies (Falke et al., 2019) show that ROUGE does not correlate well with factual consistency and we thus conduct human evaluation to better verify the improvement of CAS in factual consistency. Specifically, we randomly examine 50 constrained summaries on XSum to see whether they are improved over the unconstrained ones. Details of the human evaluation are provided in App. A. Manual analysis shows that 34% constrained summaries achieve better factual consistency, 52% are similar to their unconstrained counterparts, and only 14% become worse, which shows the effectiveness of CAS for preserving the factual consistency of abstractive summarization.

4.3 CAS with Ground-truth Constraints

We study the performance of CAS when ground-truth constraints from the reference summaries are provided. Such settings are beneficial for interactive summarization and also reveal the potential of CAS when manual (better) constraints are used.

In Tables 3 and 4, we show the performance comparison of CAS when various constraints are used, where *NER* denotes entities in the reference summary r, *miss* denotes only considering the tokens missing in the unconstrained summary s', *rand4* denotes 4 random tokens, *src* denotes only considering tokens found in the source document, *NP* denotes further adding noun phrases besides entities, and *phr4* denotes 4 continuous tokens (a phrase).

Performance Comparison on XSum. On XSum, by merely using about 4 ground-truth tokens as "hints" during inference without re-training, CAS boosts the performance of BERTSum significantly (up to 11.2 in ROUGE-2), outperforming more expensive state-of-the-art methods. Adding all constraints improves less than adding only missing constraints, possibly because the model can already generate those tokens but still "wastes" some beams to store such constraints and thus fails to search for better alternatives. Using random tokens leads to smaller gains than using entities as constraints except for ROUGE-1 and its outputs are often inarticulate, which verifies our hypothesis on the importance of named entities. The improvement of CAS is not as significant if we require the presence of constraints in the source document, which coincides with the fact that reference summaries may involve extrinsic information (Maynez et al., 2020). *phr4* brings much higher gains in ROUGE-2 and ROUGE-L since it uses a *phrase* constraint, which is favored by these two metrics.

Performance Comparison on CNN/DM. The performance gains on CNN/Daily Mail are not as significant as on XSum, possibly because the summaries in CNN/Daily Mail are much longer and

Constraint Type	XSum R-1	R-2	R-L	C
None	38.91	16.54	31.30	-
NER	46.88	20.21	33.09	**5.77**
NER-miss	46.87	21.55	33.92	4.01
rand4	46.42	17.73	33.41	4.42
rand4-miss	**52.09**	20.71	35.96	4.36
NER-miss-src	40.79	17.31	32.07	0.80
NER-NP-miss-src	43.30	18.82	33.07	2.02
phr4	47.56	27.71	38.72	4.41
phr4 (beam size 50)	49.76	**30.33**	**42.55**	4.41
BART (Lewis et al., 2019)	45.14	22.27	37.25	-
PEGASUS (Zhang et al., 2020a)	47.21	24.56	39.25	-

Table 3: **Decoding with various types of ground-truth constraints**. C denotes the average number of tokens in the constraint set $\mathcal{C}$ (after tokenization). Beam size is 5 unless otherwise specified.

Constraint Type	CNN/DM R-1	R-2	R-L
None	42.00	19.44	38.98
NER	43.31	19.57	40.05
phr4	43.37	21.57	40.82
phr4 (beam size 20)	**45.14**	**23.21**	**42.43**
BART (Lewis et al., 2019)	44.16	21.28	40.90
PEGASUS (Zhang et al., 2020a)	44.17	21.47	41.11

Table 4: **Effectiveness of CAS when ground-truth constraints are provided** on CNN/Daily Mail.

we use a similar number of constraints on both datasets. Nevertheless, CAS still helps BERTSum achieve state-of-the-art performance by only using *one* ground-truth phrase as constraint.

4.4 Performance Analysis

We argue that the improvement of CAS is not simply achieved by random insertion of tokens with the following observations.

CAS vs. Random Insertion. Different from random insertion, CAS usually inserts constraints at proper positions and often corrects unfaithful information (see Table 1 and more examples in App. C). To further study the performance gap between CAS and random insertion, we append the same constraints to the end of unconstrained summaries and conduct automatic evaluation. As listed in Table 5, randomly appending 4 to 5 ground-truth tokens cannot obtain the same performance gains as CAS.

Method	XSum R-1	XSum R-2	XSum R-L
Random (NER-miss)	47.03	20.79	34.28
CAS (NER-miss)	**48.46**	**24.17**	**37.47**
Random (phr4)	46.71	26.11	35.69
CAS (phr4)	**49.76**	**30.33**	**42.55**

Table 5: Random insertion cannot achieve performance gains as significant as CAS.

Larger beam size leads to better performance. In Fig. 1, we also observe that using larger beam size consistently leads to better performance for CAS when the same constraints are used, while the performance change is negligible for unconstrained summarization (detailed numbers in App. B). Such results further indicate that the huge gains brought by CAS are based upon a better decoding procedure that exploits model potentials through the guidance of constraints, rather than merely access to a few ground-truth tokens.

4.5 Runtime Efficiency

For automatic constraint discovery, since keyphrase extraction is independent of the summarization stage, one can use the same keyphrases on different summarization models and the total time is thus amortized. For model inference, a detailed runtime comparison between unconstrained and constrained generation is shown in Table 6. The overhead of constrained generation is acceptable when the beam size is not large. Moreover, a faster

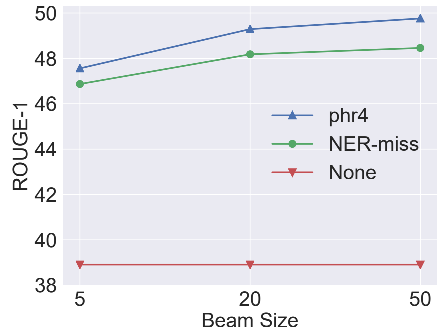

Figure 1: **Impact of beam size**. Larger beam size consistently leads to better performance for CAS.

Beam Size	Unconstrained	Constrained
5	25min	33min
10	44min	68min
20	70min	4h
50	2.5h	10h

Table 6: **Comparison of inference time** on the XSum dataset when about 4 tokens are added as constraints. Results are on one GTX 1080 Ti GPU.

DBA implementation (Hu et al., 2019b) would further reduce the runtime.

5 Conclusion and Future Work

In this paper, we propose constrained abstractive summarization to preserve the factual consistency of abstractive summarization. We demonstrate that CAS leads to higher-quality, especially more factually consistent summaries and CAS has significant benefits in both standard and interactive summarization scenarios. For future work, we will explore better constraint discovery techniques to exploit the potentials of CAS and investigate negative constraints, which ensure that constrained models never generate certain inconsistent information.

Acknowledgments

We thank Junxian He, Mike Lewis, Yanru Qu, and Jiaming Shen for helpful discussions.

References

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Shanbo Cheng, Shujian Huang, Huadong Chen, Xin-Yu Dai, and Jiajun Chen. 2016. PRIMT: A pick-revise framework for interactive machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1240–1249, San Diego, California. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Yang Gao, Christian M. Meyer, and Iryna Gurevych. 2018. APRIL: Interactively learning to summarise by combining active preference learning and reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4120–4130, Brussels, Belgium. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 166–175.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019a. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019b. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

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*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. 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.

Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han. 2020. Facet-aware evaluation for extractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4941–4957, Online. Association for Computational Linguistics.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online. Association for Computational Linguistics.

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

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.

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

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. 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 1314–1324, New Orleans, Louisiana. 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.

Ori Shapira, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2020. Evaluating interactive summarization: an expansion-based framework. *arXiv preprint arXiv:2009.08380*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325*.

Si Sun, Chenyan Xiong, Zhenghao Liu, Zhiyuan Liu, and Jie Bao. 2020. Joint keyphrase chunking and salience ranking with bert. *arXiv preprint arXiv:2004.13639*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020b. Pointer: Constrained text generation via insertion-based generative pre-training. *arXiv preprint arXiv:2005.00558*.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Boosting factual correctness of abstractive summarization with knowledge graph. *arXiv preprint arXiv:2003.08612*.

A Human Evaluation

For human evaluation, we compare the constrained summary s and unconstrained summary s' of a source document generated by the same base model for an apple-to-apple comparison. We mark CAS as *better* if it provides additional information supported by the reference summary in s or corrects factual inconsistencies in s' while keeping (most of) the original information in s', and mark CAS as *worse* if it leads to unfaithful or unsmooth s. We define *tie* if s brings information not found in the reference summary but factual (*i.e.*, supported by the source document).

B Comparison of Various Beam Sizes

As a supplement to Fig. 1, we list the detailed results of CAS when different beam sizes are used in Tables 7 and 8. The observations on XSum and CNN/Daily Mail are consistent – larger beam size leads to better performance for CAS.

Method	**XSum**			
	R-1	**R-2**	**R-L**	**B**
None	38.91	16.54	31.3	5
NER-miss	46.87	21.55	33.92	5
NER-miss	47.83	23.00	36.13	10
NER-miss	48.18	23.64	36.86	20
NER-miss	**48.46**	**24.17**	**37.47**	50
phr4	47.56	27.71	38.72	5
phr4	49.29	29.64	41.78	20
phr4	**49.76**	**30.33**	**42.55**	50

Table 7: **Impact of beam size in CAS on the XSum dataset**. B denotes the beam size. When varying the beam size of unconstrained decoding to 10/20/50, its performance change is negligible and thus unlisted.

Method	**CNN/DM**			
	R-1	**R-2**	**R-L**	**B**
None	42.00	19.44	38.98	5
NER	43.31	19.57	40.05	5
NER	**44.75**	**21.34**	**41.60**	20
phr4	43.37	21.57	40.82	5
phr4	44.77	22.80	42.14	10
phr4	**45.14**	**23.21**	**42.43**	20

Table 8: **Impact of beam size in CAS on the CNN/Daily Mail dataset**. B denotes the beam size.

C More Examples

In Table 9, we show more examples of CAS that are randomly sampled from the XSum dataset. We observe that CAS is able to correct factual inconsistencies when the major semantic of the unconstrained system summary is relevant but not so effective when its semantics is too far away from the reference summary. Such observations suggest that one could possibly obtain better results by using stronger base models such as BART (Lewis et al., 2019).

Constraint Set: ['three'] **Reference**: **three** men have been found guilty of killing a rival drug dealer in a gang-related revenge attack . **Unconstrained**: **four** men have been convicted of killing a father-of-two in a " vicious " attack in a denbighshire car park . **Constrained**: **three** men have been convicted of killing a father-of-two in a " vicious " attack in a denbighshire car park .
Constraint Set: ['monetary policy committee'] **Reference**: uk interest rates have been held at 0.5 % again by the bank of england 's **monetary policy committee** -lrb- mpc -rrb- . **Unconstrained**: the bank of england -lrb- **mpc** -rrb- has voted to keep interest rates at its current rate for the first time this year . **Constrained**: the bank of england **monetary policy committee** -lrb- mpc -rrb- has voted to keep interest rates at its current rate .
Constraint Set: ['portuguese', 'tuesday'] **Reference**: six people died and another two were missing after an explosion on **tuesday** evening destroyed a fireworks factory near the **portuguese** town of lamego . **Unconstrained**: the authorities in the **brazilian** city of **porto** have widened their search for eight people killed in an explosion at a factory . **Constrained**: the authorities in the eastern **portuguese** city of **porto** have widened their search for eight people killed in **tuesday** 's explosion at a factory .
Constraint Set: ['wales', 'public health wales'] **Reference**: an outbreak of 189 cases of measles has been reported in swansea , neath and port talbot , **public health wales** says . **Unconstrained**: the number of school children caught measles in swansea and neath port talbot has reached a record high . **Constrained**: **public health wales** has said it is concerned at the number of cases of measles in **wales** .
Constraint Set: ['sicily', 'cesare'] **Reference**: from the terrace of his winery near the baroque town of caltagirone in south-eastern **sicily** , **cesare** nicodemo surveys his fields of ripening vines - a glass of his finest spumante in hand . **Unconstrained**: it 's almost a year since the sicilian city of mauricio pounced its name to the rural mafia , but it 's not quite like it . **Constrained**: it 's almost a year since **sicily** 's **cesare** deglazio took on the streets of rome .
Constraint Set: ['northern ireland'] **Reference**: the leader of the catholic church in ireland has backed a call by amnesty international for an inquiry into mother-and-baby homes in **northern ireland** . **Unconstrained**: the archbishop of canterbury justin martin has called for an investigation into the transfer of human remains from the catholic church . **Constrained**: the archbishop of canterbury justin martin has called for an immediate investigation into the transfer of human remains to **northern ireland** .
Constraint Set: ['xinjiang'] **Reference**: china has executed eight people in the north-western region of **xinjiang** , for what it calls " terrorist " attacks , reports the state news agency xinhua . **Unconstrained**: china 's state-run news agency xinhua has executed three men for terrorism , state media report . **Constrained**: china 's state-run news agency xinhua says eight men convicted of terrorism offences in **xinjiang** have been executed .

Table 9: **Randomly sampled examples of CAS with keyphrase constraints**. We filter the keyphrases such that they are not in the unconstrained summary but in the reference to verify the effectiveness of CAS for interactive summarization.