Improving the Faithfulness of Abstractive Summarization via Entity Coverage Control

Anonymous ACL submission

Abstract

Abstractive summarization systems leveraging pre-training language models have achieved superior results on benchmark datasets. However, such models have been shown to be more prone to hallucinate facts that are unfaithful to the input context. In this paper, we propose a method to remedy entity-level extrinsic hallucinations with **E**ntity **C**overage **C**ontrol (ECC). We first compute entity coverage precision and prepend the corresponding control code for each training example, which implicitly guides the model to recognize faithfulness contents in the training phase. We further extend our method via intermediate fine-tuning on large but noisy data extracted from Wikipedia to unlock zero-shot summarization. We show that the proposed method leads to more faithful and salient abstractive summarization in supervised fine-tuning and zero-shot settings according to our experimental results on three benchmark datasets XSum, Pubmed, and SAMSum of very different domains and styles.

1 Introduction

Abstractive summarization aims to generate a compact and fluent summary that preserves the most salient content of the source document. Recent advances in pre-trained language models (Devlin et al., 2018; Liu and Lapata, 2019; Lewis et al., 2020) have led to improvements in the quality of generated summaries.

However, one prominent limitation of existing abstractive summarization systems is the lack of faithfulness of generated outputs. Faithful summaries should only contain content that can be derived from the source document instead of hallucinated or fabricated statements. Cao et al. (2018); Kryściński et al. (2019) showed that about 30% of the summaries generated by seq2seq models suffer from the hallucination phenomenon at either the entity level or the summary level. Table 1 shows an example of a model generated summary with hallucinated entities. The BBC article discusses a teenage science competition streamed on the Youtube website, while a BART-based summarizer makes up the term 'Gumtree' instead. Such hallucinations may cause factual errors and hinder the practical use of summarization models.

Source: *When the experiments are eventually run, the results will be streamed live on YouTube. Alongside Prof Hawking, the judging panel consists of [...]*
Summary: *Stephen Hawking joined the judging panel of a science competition on the internet education site* ***Gumtree****.*

Table 1: An example of model generated unfaithful summary due to entity hallucination from XSum dataset.

Faithfulness and factuality in abstractive summarization has received growing attention from the NLP community (Kryscinski et al., 2020; Goyal and Durrett, 2021; Zhu et al., 2021; Narayan et al., 2021). Recent works have attempted to address the hallucination problem at the entity level by reducing hallucinated entities during generation. Chen et al. (2021) proposed a post-processing method, which replaces the hallucinated entities in the generated outputs with the same type entities in the source document. However, it introduces additional errors to the summary and increases the intrinsic hallucination. Nan et al. (2021) proposed to address entity hallucination by filtering the training data and multi-task learning with summary-worthy named-entities classification. However, the method sacrifices part of the training data and decreases the quality of the summary.

To address the above issues, we propose to solve entity hallucination by guiding the model learning process with entity control code (ECC) (Keskar et al., 2019; He et al., 2020; Fan et al., 2017). We utilize the entity coverage precision between the training document and its reference summary as faithfulness guidance and prepend it to the corresponding document in the training phase. Then, we prepend faithful control code during inference

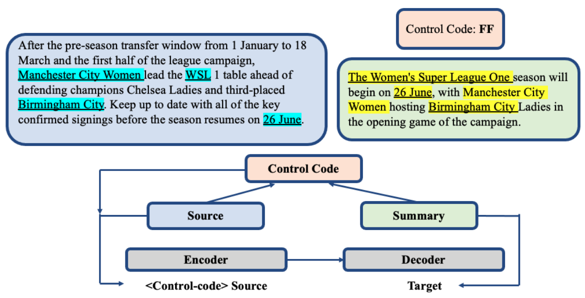

Figure 1: Entity Coverage Control for seq2seq model.

and reduce hallucinated entities effectively without decreasing the fluency and salience of generated summaries according to our experimental results. In addition, we extend control code to a Wikipedia-based intermediate tine-tuning model, which generates faithful and salient summaries across domains in the zero-shot setting. We validate our methods on three benchmark datasets across different domains, and experimental results demonstrate the effectiveness of our methods.

2 Methods

2.1 Problem Formulation

Let $D = \{(d_1, s_1), (d_2, s_2), ..., (d_n, s_n)\}$ denote a dataset composed of n document and summary pairs. During inference phase, a seq2seq model generates summary hypothesis h_i for a given document d_i by computing the probability $p_\theta(h_i|d_i)$. The generated summary h_i is expected to be faithful, which means all the information in h_i should be entailed by the source document d_i.

Following (Nan et al., 2021), we quantify entity-level hallucination with entity coverage precision $\mathbf{prec_{en}}$. It approximates the faithfulness by measuring the ratio of the named entities in the summary that are coming from the source document. Formally, it is defined as:

$$\mathbf{prec_{en}} = |\mathcal{N}(h) \cap \mathcal{N}(s)| \,/\, |\mathcal{N}(h)| \qquad (1)$$

where $\mathcal{N}(t)$ represents the set of all named entities found in a given input text t.

2.2 Entity Coverage Control

Figure 1 shows our entity coverage control method. We generate a control code C_i for each training document and reference summary pair (d_i, s_i) so the seq2seq model generates summary conditioned on both the source document d_i and its control code C_i, which is represented as $p_\theta(h_i|d_i, C_i)$.

We first compute entity coverage precision $\mathbf{prec_{en}}$ for each document and reference summary pair (d_i, s_i) in the training set D. Then, we quantize $\mathbf{prec_{en}}$ into k discrete bins, each representing a range of entity faithfulness. These bin boundaries are selected to ensure that each bin contains roughly the same number of training examples to avoid data imbalance. We then represent each bin by a special token control code C_i and add all these special tokens $\{C_1, C_2, ..., C_k\}$ to the input vocabulary of our seq2seq model.

During training, we prepend the corresponding pseudo label C_i to the input document as control code. The seq2seq model is now conditioned on both the source document d_i and its control code C_i, so it could learn different faithful level generation patterns from the control codes. Then during inference, we prepend the high faithfulness control code C_k to all documents in the test set and generate faithful summaries by $p_\theta(h_i|d_i, C_k)$.

2.3 Controllable Intermediate Fine-tuning

Large pre-trained language models (Devlin et al., 2018; Lewis et al., 2019) perform poorly in the zero-shot summarization setting since sentence salience information is not learned through pre-training tasks (Zhang et al., 2020). Thus, we propose a controllable generalized intermediate fine-tuning for zero-shot summarization.

We first generate pseudo document summary pairs from Wikipedia article dump with similar summary length (n), document length (m) and abstractiveness (a) to the target datasets following Wikitransfer (Fabbri et al., 2021). Instead of training different models for different target datasets as in WikiTransfer, we propose a unified model that generalizes well across different domains. Assume we have l target-specific pseudo training subsets $\{D_1(n_1, m_1, a_1), ..., D_l(n_l, m_l, a_l)\}$, we give each subset another special token E_i as a pseudo label to represent the target-specific pattern and also add all these special tokens $\{E_1, E_2, ..., E_l\}$ to the input vocabulary of the seq2seq model. In the training phase, we preprend the corresponding target code E_i to the document, and a summary is generated conditioned on both the source document d_i and its target control code E_i, which is represented as $p_\theta(h_i|d_i, E_i)$. This allows for control over the domain and generation style of gen-

Pubmed

Model	Entity Precision	R-1	R-2	R-L
Reference	42.85	100	100	100
BART$_{large}$	74.31	43.35	16.20	39.50
Ecc	**76.38**	**43.46**	**16.24**	**39.68**

SAMSum

Model	Entity Precision	R-1	R-2	R-L
Reference	71.20	100	100	100
BART$_{large}$	78.50	52.39	**27.89**	**43.58**
Ecc	**80.23**	**52.42**	27.69	43.34

Table 2: Experiment results in the supervised fine-tuning setting on Pubmed and SAMsum datasets, XSum results are reported in Table 3

XSum

Model	Entity Precision	FEQA	R-1	R-L
Bart	54.11	22.50	**44.78**	**36.64**
+correct	55.57	25.62	43.48	35.32
+filter	**70.49**	**26.73**	42.19	33.97
Ecc	59.38	26.51	43.82	35.97

Table 3: Performance comparison on XSum dataset.

Xsum

Model	Entity Precision	R-1	R-2	R-L
Bart	**92.61**	19.45	3.01	13.29
WikiTransfer	50.50	29.39	8.90	21.98
Ecc-zero	55.48	**30.05**	**9.72**	**22.99**

Pubmed

Model	Entity Precision	R-1	R-2	R-L
Bart	42.85	31.65	10.17	16.60
WikiTransfer	62.72	**38.64**	13.28	**19.37**
Ecc-zero	**68.13**	38.42	**13.34**	19.32

Table 4: Model performance in the zero-shot summarization setting.

Model	Faith. %	Ex. %	In. %	Quality
Bart	15.0	54.0	39.0	2.31
+correct	27.0	48.0	57.0	2.42
Ecc	28.0	**41.0**	**37.0**	**2.43**
Ecc-zero	**31.0**	48.0	38.0	1.73

Table 5: Human evaluation results of 50 test examples sampled from XSum dataset. Results with inter-annotator agreement are reported in Appendix C.

erated summaries by prepending different domain control codes during inference. The control codes are also stackable, so we can stack the target control with entity coverage control for faithful zero-shot summarization, which could be denoted as $p_\theta(h_i|d_i, C_i, E_i)$.

3 Experiments

3.1 Experiment Settings

Datasets and evaluation metric: We experiment with three mainstream datasets in different domains: news summarization dataset *XSum* (Narayan et al., 2018), scientific paper dataset *Pubmed* (Cohan et al., 2018), and dialogue summarization dataset *Samsum* (Gliwa et al., 2019). We use *ROUGE* (Lin, 2004) to measure the fluency and salience and use *Entity Precision* (Nan et al., 2021) and *FEQA* (Durmus et al., 2020) to measure the faithfulness of output summaries. We also ask expert annotators to perform a human evaluation in both summary faithfulness and quality. Implementation details are described in Appendix A.

Baselines: We compare our methods with: *BART* (Lewis et al., 2020), Bart outputs with post-processing *correction* (Chen et al., 2021), Bart with entity-based data *filtering* (Nan et al., 2021) and zero-shot Wikipedia intermediate fine-tuning *WikiTransfer* (Fabbri et al., 2021).

3.2 Automatic Evaluation

Table 2 shows the performance of our method in the supervised fine-tuning setting. Compared to the summaries generated by BART, our method increases the entity coverage precision with roughly the same summary quality. Table 3 shows the performance comparison to baselines on the XSum dataset. Our methods achieves comparable faithfulness improvements without degrading the summary quality compared to data filtering and post-processing methods.

Table 4 shows the zero-shot summarization results on XSum and Pubmed datasets. We notice BART tends to copy from the source document, so it achieves high entity coverage precision (92.61) but low summary quality. In contrast, with our intermediate fine-tuning, BART learns the characteristic of the downstream dataset and achieves a considerable improvement in ROUGE score. Compared to the baseline Wikitransfer, we see improvements in both the entity coverage precision and summary quality. Our model is also generalized cross datasets, so we use one model for different downstream targets instead of training separate models like Wikitransfer.

3.3 Human Evaluation

Table 5 shows the human evaluation results on the 50 randomly sampled subset of articles from the XSum dataset following the setting of (Chen

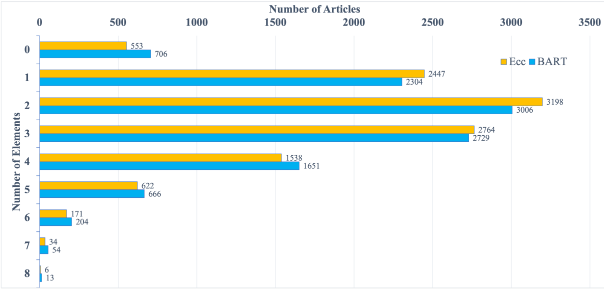

Figure 2: Number of entities in the generated summary from BART and ECC.

Model	Entity Precision	R-1	R-2	R-L
BART$_{large}$	54.11	**44.78**	21.60	36.64
LOW	51.32	44.03	21.23	36.12
MEDIUM	53.50	43.94	21.21	35.94
HIGH	**59.38**	43.82	21.15	35.97

Table 6: Comparison of summaries docoding with different control codes on XSum Dataset.

et al., 2021). Four expert annotators assign each summary output into three faithfulness categories (faithful summary, intrinsic hallucination, extrinsic hallucination) and three summary quality categories (low(1), medium (2), high(3)). Note that a summary may contain both intrinsic and extrinsic hallucinations. As the results show, our ECCmodel improves the faithfulness of the summaries without degrading summary quality, which agrees with our automatic evaluation results.

4 Analysis and Discussion

Does our model generate fewer entities to be safe? One obvious way to get higher entity coverage precision is to avoid generating entities or generating extra non-sense named entities from the source document. We show the distribution of the number of entities in the generated summaries by our model and BART in Fig 2. We see that the two distributions are very similar and have almost the same mean number of entities. As a result, we argue that our method doesn't under-generate nor over-generate entities from the source document, and we don't need to separately control the entity compression rate.

How does control code affect inference phase? We also study the effect of decoding with different control codes. We prepend different entity coverage control codes during inference on the XSum test set. As shown in Table 6, our model still generates reasonable summaries when inferred with low and medium control codes. We notice there is a trade-off between entity coverage precision and the quality of the generated summary, that summaries inferred with low control codes have higher ROUGE scores. We argue this is due to the low faithfulness level of the reference summaries in Xsum dataset (Maynez et al., 2020).

Why does BART generate hallucinated tokens? As shown in an XSum example in Table 7, fine-tuned BART generates 'Gary Anderson' according to the context 'Saints captain Anderson' , which is erroneous since the actual captain is 'Steven Anderson'. Language models contain abundant relational knowledge from pre-training data and could be extracted by masked text filling (Petroni et al., 2019). Similarly, we insert a mask token before 'Anderson' and probe untuned BART to fill the masked tokens. BART generates 'Paul Anderson' (actor) when only given the first sentence context. When given the whole news article, BART learns the context is sports-related and generates famous athletes 'Craig Anderson' (hockey athlete) and 'Gary Anderson' (football athlete) according to its pre-trained prior knowledge. The ground truth 'Steven Anderson' appears much less frequent during pre-training, so BART has a low probability of generating it correctly. We observe the same for ground truth 'Rob Kiernan', which probably appears less frequently in BART's pre-training corpus.

Document: Saints captain **<mask>** Anderson claims he was punched by Kiernan during last week's 1-1 draw between the sides. [...] **Bart:** St Johnstone's ***Gary*** Anderson says Rangers midfielder ***John*** Kiernan should face a Scottish FA disciplinary hearing over an alleged punch.
Reconstructed <mask> from 1st sentence context: Top-5: ['Paul', 'Mark', 'Tom', 'James', 'Ryan'] **Reconstructed <mask>** from full source context: Top-5: ['Craig', '***Gary*** ', 'Kier', '***Steven***', 'Anderson']

Table 7: An example of hallucinated entity analysis with mask token refilling by BART. The ground truth is 'Steven Anderson' according to web search.

5 Conclusion

In this paper, we study entity coverage control as a method to address extrinsic hallucination in abstractive summarization in both supervised and zero-shot settings. Our extensive experiment results demonstrate that our proposed method effectively reduces entity hallucination without hurting the quality of the generated summaries.

References

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

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. *arXiv preprint arXiv:2104.09061*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

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

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online. Association for Computational Linguistics.

Angela Fan, David Grangier, and Michael Auli. 2017. Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

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

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

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.

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*.

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, Online. Association for Computational Linguistics.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1808.08745*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simoes, and Ryan McDonald. 2021. Planning with entity chains for abstractive summarization. *arXiv preprint arXiv:2104.07606*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. 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*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

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 Implementation Details

We use Huggingface libraries (Wolf et al., 2020) for all our experiment implementations. Our backbone abstractive summarization model is BART-large (Lewis et al., 2020), a pre-trained denoising autoencoder language model with $336M$ parameters based on the sequence-to-sequence transformer (Vaswani et al., 2017). For fair comparison, we fine-tune BART-large on each dataset for on 8 Tesla A100 GPU pods with same learning rate $5e-5$ with weight decay using Adam optimizer (Kingma and Ba, 2014).

For entity recognition, we use a neural Named Entity Recognition (NER) system from the Stanza NLP toolkit (Qi et al., 2020) trained on the OntoNotes corpus (Weischedel et al., 2011) except for Pubmed dataset. Since Pubmed is a medical scientific article collection, we use biomedical, scientific, and clinical text Named Entity Recognition toolkit scispaCy (Neumann et al., 2019) instead.

B Representative Examples Analysis

In Table 8, we provide several representative examples from XSum dataset. Example 1 (first row) shows how our entity control method gets rid of hallucination terms from BART output. The reference summary here is not faithful since 'Los Angeles' is not covered in the source document. The correction baseline changes 'Los Angeles' to 'Mexico', which is a factual error. In contrast, the ECCoutput is totally faithful to the source document and contains salient information.

Example 2 (second row) shows the outputs decoded with different control codes during inference. We can see the output decoded with low faithfulness control code is still fluent and reasonable, but contains less faithful entities compared to the output decoded with high faithfulness control code.

Example 3 (third row) shows an example of factual statement, which is verifiable in the real world independent of the source text. The reference summary uses 'most of Wales' to summarize the county names in the source document. This type of hallucination needs more external knowledge and commonsense reasoning to decide its factuality. Our method only focuses on entity level hallucination problems instead.

C Human Evaluation Confidence

Our human evaluation follows the setting of prior work (Chen et al., 2021). We calculate the inter-

Bart: A video game based on one of the world's most popular wrestling traditions has been launched at the E3 gaming show in Los Angeles.' **Correction:** A video game based on one of the world's most popular wrestling traditions has been launched at the E3 gaming show in Mexico. **ECC:** A video game dedicated to Mexican wrestling has been released at E3. **Reference:** One of the more unusual titles at E3, the worlds largest video games exhibition held each year in Los Angeles, is Konami's Lucha Libre AAA: Heroes del Ring.
Bart: Tourists in Spain have been accused of harassing a dolphin after it became stranded on a beach. **Low Code:** A dolphin that became stranded in the sea off the coast of Spain has been harassed by a group of tourists. **High Code:** A dolphin that became stranded in the sea off the coast of Andalucia has been harassed by tourists. **Reference:** A baby dolphin has died after it was surrounded by tourists looking to take photographs on a beach in southern Spain.
Document: The warning begins at 22:00 GMT on Saturday and ends at 10:00 on Sunday. The ice could lead to difficult driving conditions on untreated roads and slippery conditions on pavements, the weather service warned. Only the southernmost counties and parts of the most westerly counties are expected to escape. Counties expected to be affected are Carmarthenshire, Powys, Ceredigion, Pembrokeshire, Denbighshire, Gwynedd, Wrexham, Conwy, Flintshire, Anglesey, ..., Rhondda Cynon Taff and Torfaen. **Reference:**The Met Office has issued a yellow weather warning for ice across most of Wales.

Table 8: Representative examples from the XSum test set.

Model	Faith. %	Ex. %	In. %	Quality
BART	15.0 ± 7.4	54.0 ± 11.2	39.0 ± 5.8	2.31 ± 0.14
ECC	28.0 ± 6.2	41.0 ± 7.2	37.0 ± 8.3	2.43 ± 0.17
ECC-zero	31.0 ± 2.8	48.0 ± 9.3	38.0 ± 7.2	1.73 ± 0.07

Table 9: Human evaluation results of 50 test examples sampled from XSum dataset.

annotator agreement with additional annotations from two other experts. We estimate the adjusted mean and 95% confidence interval from the mean and standard deviation. The full results are shown in Table 9.