Factual Error Correction for Abstractive Summarization Models

Meng Cao   **Yue Dong**   **Jiapeng Wu**   **Jackie Chi Kit Cheung**

School of Computer Science, McGill University, Montreal, QC, Canada
MILA, Montreal, QC, Canada

{meng.cao@mail, yue.dong2@mail, jiapeng.wu@mail, jcheung@cs}.mcgill.ca

Abstract

Neural abstractive summarization systems have achieved promising progress, thanks to the availability of large-scale datasets and models pre-trained with self-supervised methods. However, ensuring the factual consistency of the generated summaries for abstractive summarization systems is a challenge. We propose a post-editing corrector module to address this issue by identifying and correcting factual errors in generated summaries. The neural corrector model is pre-trained on artificial examples that are created by applying a series of heuristic transformations on reference summaries. These transformations are inspired by an error analysis of state-of-the-art summarization model outputs. Experimental results show that our model is able to correct factual errors in summaries generated by other neural summarization models and outperforms previous models on factual consistency evaluation on the CNN/DailyMail dataset. We also find that transferring from artificial error correction to downstream settings is still very challenging[1].

1 Introduction

Self-supervised methods have achieved success in a wide range of NLP tasks, and automatic summarization is no exception (Liu and Lapata, 2019; Lewis et al., 2019; Zhang et al., 2019a; Shi et al., 2019; Fabbri et al., 2019). These state-of-the-art abstractive summarization models typically finetune pre-trained transformer-based models on a summarization dataset (Vaswani et al., 2017). Despite significant improvements over previous methods in terms of automatic evaluation scores such as ROUGE (Lin, 2004), ensuring factual consistency of the generated summary with respect to the source remains challenging. For example, Cao

Source:
Jerusalem (CNN)The flame of remembrance burns in Jerusalem, and a song of memory haunts Valerie Braham as it never has before. This year, Israel's Memorial Day commemoration is for bereaved family members such as Braham. "Now I truly understand everyone who has lost a loved one," Braham said. (...)
Original: **France's** memorial day commemoration is for bereaved family members as braham. *(inconsistent)*
After Correction: **Israel's** memorial day commemoration is for bereaved family members as braham. *(consistent)*

Table 1: An example of an inconsistent system-generated summary and the output summary from our correction model. In this case, "France" is successfully corrected as "Israel".

et al. (2018) claims that about 30% of summaries generated by abstractive models contain factual errors, which greatly limits their practicality.

Different approaches have been proposed to detect or ensure the factual consistency of generated summaries, including using fact extraction or applying attention on fact triples (Cao et al., 2018; Zhang et al., 2019b; Goodrich et al., 2019), applying natural language inference or question answering models for consistency checking (Falke et al., 2019; Li et al., 2018; Wang et al., 2020) and training the model on artificial datasets (Kryściński et al., 2019). Most of these approaches either require a high-quality fact extraction model or they only focus on factual consistency *evaluation*. Improving factuality *correction* by editing inconsistent parts in generated summaries is a direction that has not been explored much.

In this work, we propose a model to improve the factual consistency of system summaries with *post-editing correction* (Table 1). Our model takes a draft summary that is generated by an abstractive summarization model and produces a corrected final summary, conditioned on the source document. In addition, our trained corrector can be

---

[1]Our data and code is available at https://github.com/mcao610/Factual-Error-Correction

used as an evaluation model for factual consistency of abstractive summaries, with the assumption that a generated summary is inconsistent if our corrector decides to make edits. To teach the model to correct errors, we train it with artificial data that has factual errors introduced using heuristics proposed by Kryściński et al. (2019).

The empirical results based on automatic and human evaluations indicate that our model not only corrects factual errors in summaries, it is also a better and more reliable factuality evaluation model than FactCC (Kryściński et al., 2019), a recent proposed factuality evaluation method. In a downstream setting where we apply the corrector to the output of an abstractive summarizer, we find that our corrector is able to accurately correct errors in the generated summaries. However, the overall recall on correcting factual errors in real system summaries remains low, suggesting the errors introduced by heuristics have a different distribution than errors made by abstractive summarization systems.

2 Background and Related Work

Previous work on factual consistency in abstractive summarization can be divided into two categories: abstractive summarization models tailored towards factual consistency (Cao et al., 2018; Zhang et al., 2019b; Li et al., 2018), and evaluation models for factual consistency in abstractive summarization (Goodrich et al., 2019; Falke et al., 2019; Kryściński et al., 2019; Wang et al., 2020).

Cao et al. (2018) proposed a dual attention module in an abstractive summarizer that attends to both the source document and to relation triples extracted from the document. Zhang et al. (2019b) propose to improve their abstractive summarization model by optimizing fact scores defined in radiology reports with reinforcement learning methods. Li et al. (2018) jointly train their model's encoder on summarization and NLI tasks. Guo et al. (2018) train an abstractive summarization system with the auxiliary tasks of question and entailment generation and show that their generated summaries are less likely to produce extraneous facts. Kumar and Cheung (2019) show that neural abstractive summarizers often assign higher posterior likelihood to perturbed contrastive summaries that are inconsistent with the source text than to human-written gold-standard ones. Concurrently to our work, Zhu et al. (2020) recently proposed

a fact-aware summarization model that uses a knowledge graph. They use a pre-trained corrector module to modify generated summaries. Concurrent to our work, Dong et al. (2020) proposes factual correction models that leverages knowledge learned from question answering models via span selection. Their models employ single or multi-masking strategies to either iteratively or auto-regressively replace entities.

In terms of evaluating abstractive summarization models for factual consistency, Goodrich et al. (2019) proposed a metric to check factual consistency by checking the overlapped fact triples between a source document and generated text on Wikidata. Falke et al. (2019) shows that factual error detection is a difficult task on its own and adapting entailment models for factual error detection do not offer the desired performance. Kryściński et al. (2019) finetune a BERT model on heuristically-created data with six types of rule-based text transformations for factual consistency checking. Wang et al. (2020) propose a framework for measuring inconsistencies in abstractive summarization by answering questions based on both generated summaries and documents.

3 Proposed Approach

In this section, we describe our procedure of introducing artificial errors in the datasets for training and propose our end-to-end error corrector model.

3.1 Dataset of Artificial Corruptions

Inspired by a recent study of error types made by state-of-the-art summarization system, we artificially created a weakly-supervised training dataset based on the text transformations proposed by Kryściński et al. (2019).

Given a source text d and the reference summary s, we corrupt the reference summary into an inconsistent summary s' with a randomly sampled corruption rule (described below) with probability α; otherwise, we keep $s' = s$ with probability $1 - \alpha$. We set $\alpha = 0.3$ to match the factuality error rate in real abstract summaries based on a recent study (Cao et al., 2018). The training data consists of triplets (s', s, d).

Error Corruptions Four types of errors are used to create the inconsistent summaries: *Entity*, *Number*, *Date*, and *Pronoun* errors. They are the most common types of errors in abstractive summaries based on our manual inspection of

	Overall Acc.	Consistency checking		
		Prec.	Recall	F1
Corrupted	84.38%	0.79	0.95	0.86
Clean		0.93	0.74	0.82

Table 3: Performance of our model on consistency checking on our test set of artificial corruptions. Corrupted and clean refer to the subsets of the test set that were artificially perturbed or not perturbed, respectively.

Source:
(CNN) Gastrointestinal illness has gripped 100 people on the cruise ship Celebrity Infinity, according to a report from the Centers for Disease Control. Of the ship's 2,117 passengers, 95 have suffered from vomiting, diarrhea and other symptoms, the CDC said. (...)

Reference Summary:
100 passengers and crew members have been sickened on Celebrity Infinity. The ship, which is based on the West Coast, left San Diego in late March .

Corrupted Summary:
95 passengers and crew members have been sickened on Celebrity Infinity. The ship, which is based on the West Coast, left San Diego in late March .

Table 2: An example of a *Number* corruption in the training set. The incorrect number "95" also appears in the source document.

100 abstractive system-generated summaries that are sampled from the dataset of Kryściński et al. (2019) (henceforth, the K2019 dataset). Unlike Kryściński et al. (2019), we corrupt the reference summary rather than sentences sampled from the source document.

In the first four types of error constructions, we utilize a swapping strategy to introduce errors. For *Entity*, *Number*, and *Date* swapping, one entity in the reference summary is selected and swapped with another random entity of the same type[2] in the source document. For *Pronoun* swapping, one pronoun was extracted and swapped with another one of a matching syntactic case. Table 2 shows one example of a corruption.

3.2 Training Objective and Models

With the artificial training data consisting of triplets (s', s, d), the goal of the corrector is to generate the correct summary s based on the inconsistent summary s' and the source d. This can be expressed as a problem of maximizing the likelihood of $P(s|s', d)$ in an encoder-decoder model. We concatenate s' and d as input to the encoder (s' and d are separated by a separation token) and train the decoder to generate s.

We use BART (Lewis et al., 2019) as the basis of our summary corrector because of its demonstrated level of performance on conditional text generation tasks. BART is a sequence-to-sequence auto-regressive transformer model that is pre-trained as a denoising auto-encoder. One appealing aspect about BART is that it is pre-trained on

---

[2]All the entities are extracted using a pre-trained NER model in spaCy https://spacy.io/.

a denoising task. Specifically, given an input sentence that is corrupted by text infilling, token deletion as well as other text transformations, BART is trained to output the original sentence. This pre-training task is similar to our summary correction task in which we can regard the corrupted or generated summary as the noisy input and in this case the noise is the inconsistent content in the summary.

4 Experiments

4.1 Evaluation Tasks and Measures

We evaluate our model on two tasks: factual consistency checking and error correction.

Factual consistency checking For this task, the model needs to classify each original input summary as *consistent* or *inconsistent* with respect to the source text. It is thus a binary classification task for which we report **accuracy**, as well as **precision**, **recall**, and **F1**.

We interpret the output of our corrector model as a classification decision as follows. If the corrector makes any change to the original input summary, we consider this to be a prediction of the *inconsistent* class. Otherwise, the corrector makes no change and we consider this a prediction of the *consistent* class.

Error correction For this task, the model must correct inconsistencies in the original summary (in any) with respect to the source text.

We define **correction accuracy** as the proportion of original summaries that are correctly changed by our corrector. On our artificial test set, an input summary is considered successfully corrected if the corrected summary matches the reference summary exactly. For the K2019 dataset, no reference corrections are available. We instead

conducted a human evaluation to check the consistency of the corrected output. We read the original and corrected summaries as well as the source document to determine whether a summary is successfully corrected by our model.

4.2 Datasets

We use two datasets for our experiments. The first is the dataset of artificial corruptions described in Section 3.1, which we create by taking samples from the CNN/DailyMail dataset. There are in total 287,227 samples in the training set, and we corrupted 30% of them (85,583). This results in 16,858/35,113/13,408/20,204 date/entity/number/pronoun corrupted samples respectively. We refer the other 201,644 training samples as clean samples. We also create artificial validation and test set for model selection and evaluation. In the test set, there are 5,780 corrupted samples and 5,710 clean samples.

The second dataset we use is the K2019 test set of Kryściński et al. (2019). This dataset contains 503 summaries generated by different recent neural abstractive summarizers, which have been manually labeled for whether they contain an inconsistency.

We evaluate our model on both datasets. We did not use baselines for the artificial test set since it is simply used as a check to demonstrate our model's performance in the artificial setting. The more meaningful evaluations are on K2019 consistency checking and error correction.

4.3 Corrector Training Details

We use the BART implementation from fairseq as the basis of our corrector.[3] The pre-trained BART model is fine-tuned on our training dataset for 10 epochs as described in Section 3.2. The learning rate is set to 3e-5. All our experiments is done on 4 NVIDIA Tesla V100 GPUs. The training process takes about 12 hours.

5 Results

Artificial corruptions Table 3 shows the consistency checking performance of our corrector model on our artificial test set. The high classification accuracy and F1 scores indicate that our model is able to identify these artificially injected errors.

---

[3] https://github.com/pytorch/fairseq/blob/master/examples/bart

Model	Accuracy	Macro F1
BERT+MNLI	51.51%	0.0882
BERT+FEVER	52.07%	0.0857
FactCC	74.15%	0.5106
FactCCX	72.88%	0.5005
Our model	**83.10%**	**0.6121**

Table 4: Factual consistency checking performance on the K2019 test set. Macro F1 is the macro-average of the F1 between the two classes.

For error correction, among the 5780 corrupted summaries in the test set, 62.13% are corrected by the model to exactly match the reference summary. For the 5710 clean summaries, the model made changes to 26.27% of them, which results in 73.73% correction accuracy on clean summaries. These results show that the model is able to correct majority of the test samples even under our strict evaluation measure.

K2019 Table 4 shows the consistency checking results on the K2019 test set. Compared with FactCC, our model improves performance by more than 10% higher in accuracy and 0.11 in macro F1-score. This result is interesting considering that the model is trained on a generation task rather than on classification.

As for correction performance, Table 6 shows the evaluation result of our human evaluation. Among 62 inconsistent summaries in the test set, the corrector model made changes to 19 summaries, of which 11 were successfully corrected and 7 remained inconsistent. For the remaining 441 consistent summaries in the test set, changes are made to 39 summaries and the model changed the meaning of 5 samples. In conclusion, with 17.74% probability that our model can successfully correct an inconsistent summary and 1.13% probability that it will corrupt a consistent one. Compared with the correction rate of 62.13% on the artificial test set, much lower correction rate on the real test set suggests that there is still a gap between the two settings. The error types in the training set are not able to represent the diverse errors made by summarization systems.

Output Analysis Table 5 shows several input and output summaries of our corrector model together with the source document fragments. In the

Article: Jerusalem (CNN)The flame of remembrance burns in Jerusalem, and a song of memory haunts Valerie Braham as it never has before. (...) "Now I truly understand everyone who has lost a loved one," Braham said. Her husband, Philippe Braham, was one of 17 people killed in January's terror attacks in Paris. He was in a kosher supermarket when a gunman stormed in, killing four people, all of them Jewish. (...)			

Article: Jerusalem (CNN)The flame of remembrance burns in Jerusalem, and a song of memory haunts Valerie Braham as it never has before. (...) "Now I truly understand everyone who has lost a loved one," Braham said. Her husband, Philippe Braham, was one of 17 people killed in January's terror attacks in Paris. He was in a kosher supermarket when a gunman stormed in, killing four people, all of them Jewish. (...)

Original: **Valerie** braham was one of 17 people killed in january's terror attacks in paris. *(inconsistent)*

Corrected: **Philippe** braham was one of 17 people killed in january's terror attacks in paris. *(consistent)*

Article: (...) Thursday's attack by al-Shabaab militants killed 147 people, including 142 students, three security officers and two university security personnel. The attack left 104 people injured, including 19 who are in critical condition, Nkaissery said. (...)

Original: **147** people, including 142 students, are in critical condition. *(inconsistent)*

Corrected: **19** people, including 142 students, are in critical condition. *(inconsistent)*

Article: (CNN) Officer Michael Slager's five-year career with the North Charleston Police Department in South Carolina ended after he resorted to deadly force following a routine traffic stop. (...) His back is to Slager, who, from a few yards away, raises his gun and fires. Slager is now charged with murder. The FBI is involved in the investigation of the slaying of the father of four. (...)

Original: **Slager** is now charged with murder. *(consistent)*

Corrected: **Michael Slager** is now charged with murder. *(consistent)*

Article: (CNN)The announcement this year of a new, original Dr. Seuss book sent a wave of nostalgic giddiness across Twitter, and months before publication, the number of pre-orders for "What Pet Should I Get?" continues to climb. (...) It features the spirited siblings from the beloved classic "One Fish Two Fish Red Fish Blue Fish" and is believed to have been written between 1958 and 1962. (...)

Original: **Seuss** book sent a wave of nostalgic giddiness across twitter. *(consistent)*

Corrected: **"One Fish Two Fish Red Fish Blue Fish"** book sent a wave of nostalgic giddiness across twitter. *(inconsistent)*

Table 5: Examples of applying our corrector to the output of a summarizer. In the first example, the original inconsistent summary is successfully corrected. In the second example, the summary remains false after correction. The third and fourth examples show changes made on consistent summaries. Colored content in articles are support for summaries.

Input	# Samples	After Correction	
		cons.	*incons.*
consistent	441	436	5
inconsistent	62	11	51

Table 6: Human evaluation results on error correction on the K2019 dataset.

second example, the model correctly replaced 147 with 19, but was not able to correctly remove "including 142 students", which is a larger modification to the original summary. More examples can be found in the Appendix.

6 Conclusions

In this paper, we proposed a novel approach to correct inconsistent content in summaries generated by abstractive summarization models. We train an end-to-end correction model with artificial examples created by corrupting reference summaries. Our model achieved promising performance on our artificial test set and outperformed previous models on the manually annotated test set by wide margins. Our human evaluation indicates that our model is able to correct some factually inconsistent summaries generated by abstractive summa-

rization model. However, low recall on the inconsistent summaries and false positive samples remain as challenges.

References

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

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Tobias Falke, Leonardo FR 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.

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.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia. Association for Computational Linguistics.

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

Krtin Kumar and Jackie Chi Kit Cheung. 2019. Understanding the Behaviour of Neural Abstractive Summarizers using Contrastive Examples. 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 3949–3954, Minneapolis, Minnesota. 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*.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441.

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.

Tian Shi, Ping Wang, and Chandan K. Reddy. 2019. LeafNATS: An open-source toolkit and live demo system for neural abstractive text summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 66–71, Minneapolis, Minnesota. 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.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D Manning, and Curtis P Langlotz. 2019b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541*.

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 Appendix

A.1 Summary Correction Examples

Table 5 shows examples of generated summaries and outputs from our corrector. We put more examples here:

1. *Article*: (CNN)In case you needed a reminder that President Barack Obama isn't running for office again, he just alienated not only Republicans, who have largely resented him from day one, but the progressive base of Democratic voters. Obama has argued with the progressive potentate Elizabeth Warren, calling her "wrong" on trade policy. (...)
   Original: **Obama** has argued with the progressive potentate elizabeth warren. *(consistent)*
   Corrected: **President Obama** has argued with the progressive potentate elizabeth warren. *(consistent)*

2. *Article*: (The Hollywood Reporter)The author of a 2006 novel has accused the "Avengers" director and "Cabin" director Drew Goddard of stealing his idea. (...) Gallagher is basing his claim on the works' similar premises: Both feature a group of young people terrorized by monsters while staying at a cabin in what is revealed to be (spoiler alert) a horror-film scenario designed by mysterious operators. (...)
   Original: **Gallagher** is basing his claim on the works' names and personalities. *(inconsistent)*
   Corrected: **Peter Gallagher** is basing his claim on the works' names and personalities. *(inconsistent)*

3. *Article*: (CNN)Too little, too late. (...) After the story of the statue caught fire online this weekend, Poulin publicly apologized Monday for his "most unsettling sculpture" in a letter to The Hollywood Reporter. (...)
   Original: **Poulin** publicly apologized for his "most unsettling sculpture" in a letter to the hollywood reporter. *(consistent)*
   Corrected: **Dave Poulin** publicly apologized for his "most unsettling sculpture" in a letter to the hollywood reporter. *(consistent)*

4. *Article*: (CNN)If I had to describe the U.S.-Iranian relationship in one word it would be "overmatched". (...) America is alienating some of our closest allies because of the Iran deal, and Iran is picking up new ones and bolstering relations with old ones who are growing more dependent because they see Iran's power rising. (...)
   Original: **Iran** is alienating some of our closest allies because of the iran deal, and iran is picking up new ones. *(inconsistent)*
   Corrected: **America** is alienating some of our closest allies because of the Iran deal, and Iran is picking up new ones. *(consistent)*

5. *Article*: (...) McHenry quickly issued an apology, blaming the incident on a moment of intense frustration but admitting her mistake and accepting responsibility. (...)
   Original: **Mchenry** apologizes to the incident on a moment of intense frustration. *(consistent)*
   Corrected: **Britt Mchenry** apologizes to the incident on a moment of intense frustration. *(consistent)*

6. *Article*: Boston (CNN)When the bomb went off, Steve Woolfenden thought he was still standing. That was because, as he lay on the ground, he was still holding the handles of his son's stroller. He pulled back the stroller's cover and saw that his son, Leo, 3, was conscious but bleeding from the left side of his head. (...)
   Original: **Steve woolfenden**, 3, was conscious but bleeding from the left side of his head. *(inconsistent)*
   Corrected: **Leo Woolfenden**, 3, was conscious but bleeding from the left side of his head. *(consistent)*

7. *Article*: (CNN)Mercedes driver and F1 championship leader Lewis Hamilton stole pole position for Sunday's Chinese Grand Prix from teammate and fierce rival Nico Rosberg in dramatic fashion. (...) He did, however, find time to congratulate fellow German driver Sebastian Vettel, who will start in third after the Ferrari driver surprisingly won the Malaysian GP two weeks ago. (...)
   Original: **Sebastian vettel** won the malaysian gp two weeks ago. *(consistent)*

Corrected: **Nico Rosberg** won the malaysian gp two weeks ago. *(inconsistent)*

8. *Article*: (CNN)At least 21 people were killed during a shipwreck off the northern coast of Haiti, the country's civil protection directorate told CNN on Thursday. (...) So far, 11 victims – eight men and three women – have been identified, Celestin said. (...)
Original: The 21 people are **three** women and three women have been identified. *(inconsistent)* The 21 people are **eight** women and three women have been identified.
Corrected: *(inconsistent)*

9. *Article*: (CNN)Oklahoma Gov. Mary Fallin signed a bill on Friday that would allow the state to perform executions with nitrogen gas if lethal injection is ruled unconstitutional or becomes unavailable. Nitrogen causes a quick loss of consciousness and then death from lack of oxygen, Fallin's office said in a press release. (...)
Original: **Nitrogen** causes a quick loss of consciousness and then death from lack of oxygen. *(consistent)*
Corrected: **Netherlands** causes a quick loss of consciousness and then death from lack of oxygen. *(inconsistent)*