NumHG: A Dataset for Number-Focused Headline Generation

Jian-Tao Huang[†], Chung-Chi Chen[‡], Hen-Hsen Huang[§], Hsin-Hsi Chen[†]
[†]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[‡]AIST, Japan
[§]Institute of Information Science, Academia Sinica, Taiwan
jthuang@nlg.csie.ntu.edu.tw, c.c.chen@acm.org,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

Abstract

Headline generation, a key task in abstractive summarization, strives to condense a full-length article into a succinct, single line of text. Notably, while contemporary encoder-decoder models excel based on the ROUGE metric, they often falter when it comes to the precise generation of numerals in headlines. We identify the lack of datasets providing fine-grained annotations for accurate numeral generation as a major roadblock. To address this, we introduce a new dataset, the NumHG, and provide over 27,000 annotated numeral-rich news articles for detailed investigation. Further, we evaluate five well-performing models from previous headline-generation tasks using human evaluation in terms of numerical accuracy, reasonableness, and readability. Our study reveals a need for improvement in numerical accuracy, demonstrating the potential of the NumHG dataset to drive progress in number-focused headline generation and stimulate further discussions in numeral-focused text generation.

Keywords: Headline generation, news headline, numeracy

1. Introduction

The pursuit of headline generation is an endeavor to distill the essential elements of an article into a single line of text. Though related, this task poses a more significant challenge than merely extracting sentences for summarization, as it requires crafting a new sentence encapsulating the same core ideas. As Matsumaru et al. (2020) have demonstrated, the performance of state-of-the-art encoder-decoder models, as judged by the ROUGE metric, is commendable. However, these models sometimes falter by creating inappropriate headlines. The crux of the issue lies in selecting words that, although superficially similar to the source text, may misrepresent the meaning and be unconnected to the original article. A critical observation from our research is that inaccuracies in using ”numerals” are a pivotal factor contributing to these erroneous headlines.

Despite this, datasets that offer fine-grained annotations and frameworks for accurate numeral generation in news headlines are in short supply. In response to this deficit, we propose a novel dataset designed to explore this issue comprehensively. Table 1 demonstrates an example from our proposed dataset. Our objective is to ensure accurate numeral generation in headlines, and as such, we provide detailed annotations on how to secure the correct numeral through specific operations. As no existing public datasets align with our task's unique characteristics, we annotated more than 27,000 numeral-rich news articles to further probe this research direction. These extensive annotations enable us to identify several unique characteristics of numerals in news headlines, thereby distinguishing our task settings from those of current numeral-related datasets.

We evaluate five models (Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020a; Wang et al., 2022; Liu et al., 2022) previously shown to perform well in headline generation tasks, conducting a human evaluation across three dimensions: numerical accuracy, reasonableness, and readability. Our findings suggest that alongside the traditional focuses of reasonableness and readability, there remains significant room for improvement in numerical accuracy. Through the release of our proposed NumHG dataset, we hope to accelerate progress in number-centric headline generation and stimulate further discussion on numeral-focused text generation.

2. Related Work

The task of headline generation, a form of text summarization, endeavors to condense a lengthy source text into a succinct summary. Text summarization approaches typically fall into two categories: extractive and abstractive. Extractive approaches involve selecting fitting sentences from the source text to serve as the summary, while abstractive approaches strive to create new sentences to encapsulate the source text. The concept of headline generation aligns more closely with abstractive methodologies.

The emergence and development of large-scale pre-trained models (Raffel et al., 2020; Lewis et al., 2020; Zhang et al., 2020a) have notably advanced the capabilities of abstractive summarization models to the extent that they now outperform extractive models. Some recent studies (Dou et al.,

News: At least **30** gunmen burst into a drug rehabilitation center in a Mexican border state capital and opened fire, killing **19** men and wounding **four** people, police said. Gunmen also killed **16** people in another drug-plagued northern city. The killings in Chihuahua city and in Ciudad Madero marked one of the bloodiest weeks ever in Mexico and came just weeks after authorities discovered **55** bodies in an abandoned silver mine, presumably victims of the country's drug violence. More than **60** people have died in mass shootings at rehab clinics in a little less than **two** years. Police have said **two** of Mexico's **six** major drug cartels are exploiting the centers to recruit hit men and drug smugglers, ...
Headline (Question): Mexico Gunmen Kill ____
Answer: 35
Annotation: Add(19,16)

Table 1: An annotation example in NumHG.

Operator	Description	Ratio
Copy(v)	Copy v from the article	65.00%
Trans(e)	Covert e into a number	17.37%
Paraphrase(v_0,n)	Paraphrase the form of digits to other representations	8.27%
Round(v_0,c)	Hold c digits after the decimal point of v_0	3.10%
Subtract(v_0,v_1)	Subtract v_1 from v_0	2.15%
Add(v_0,v_1)	Add v_0 and v_1	1.73%
Span(s)	Select a span from the article	1.34%
Divide(v_0,v_1)	Divide v_0 by v_1	0.54%
Multiply(v_0,v_1)	Multiply v_0 and v_1	0.50%

Table 2: Overview of predefined operators. v, v_0 and v_1 denote the selected numerals, and e denotes the English word. s and c denote a span from the article and a constant, respectively.

Corpus	# Sents	# Words	# Nums
Dolphin18K (Huang et al., 2016)	2.6	30.6	4.4
AQUA-RAT (Ling et al., 2017)	2.2	32.5	4.2
Math23K (Wang et al., 2017)	1.6	28.0	3.1
MathQA (Amini et al., 2019)	2.0	37.9	4.5
SVAMP (Patel et al., 2021)	2.8	31.8	3.2
NumHG (Proposed)	**9.4**	**191.8**	**13.7**

Table 3: Comparison of different corpora.

2021; Wang et al., 2022; Liu et al., 2022) emphasize the significance of keyword sentences, asserting that these should be leveraged as guides for summary generation. GSum (Dou et al., 2021), for example, initially performs extractive summarization, then incorporates the extractive summaries into the input for abstractive summarization. Despite experimental evidence supporting GSum's effectiveness, Wang et al. (Wang et al., 2022) argue that extractive summaries do not provide a reliable or flexible guide, potentially leading to information loss or noisy signals.

To tackle this issue, Season (Wang et al., 2022) adopts a dual approach, learning to predict the informativeness of each sentence and using this predicted information to guide abstractive summarization. Meanwhile, BRIO (Liu et al., 2022) employs pre-trained abstractive models to generate candidate summaries, assigning each a probability mass according to their quality and defining a contrastive loss across the candidates. By considering both token-level prediction accuracy and sequence-level coordination, BRIO combines cross-entropy loss and contrastive loss for abstractive summarization.

Notably, the majority of these works focuses on the selection of words and the structure of sentences. However, our work diverges significantly as it specifically tackles the problem of numeral accuracy in headline generation—a factor only discussed in a few studies (Chu et al., 2020; Chen et al., 2021). Our newly proposed NumHG dataset, comprising over 27,000 annotated numeral-rich news articles, provides a valuable resource for enhancing the performance of numeral-aware headline generation tasks.

3. Dataset

3.1. Dataset Design

This section provides a comprehensive introduction to the proposed NumHG dataset. Following(Fabbri et al., 2019), the primary source of our news articles is Newser[1], a news aggregation platform that curates top stories from numerous U.S. and international outlets. Articles on Newser typically contain approximately 200 to 300 words. Our focus for the NumHG dataset is news articles with numeral-infused headlines. Consequently, we eliminate articles without numerals in the headline. As a further restriction, NumHG is centered on headlines featuring only a single number, leading us to exclude articles with more than one numeral in the headline. These filtering processes result in a dataset of 27,746 instances.

For accurate numeral generation in headlines, the model may need to manipulate the numerals in the article body or perform basic calculations. For instance, the headline numeral in the example provided in Table 1 requires a simple calculation. Given the absence of suitable existing datasets for this purpose, we devise an annotation scheme to understand the operations between numerals in the news articles and the headlines. After sampling 3,000 instances for operator distribution analysis, we define a set of operators for our annotation guideline, as shown in Table 2.

Note that, for these numerals written in different

[1] https://www.newser.com/

formats, we designed the “Trans” operator to convert them into a uniform format (e.g., “Trans(two) => 2”, “Trans(second) => 2”, “Trans(dozen) => 12”). Therefore, annotators had to use predefined operators and numerals directly appearing in the news text (including these different formats) to pass automated validation. Automated assessment required annotators to further judge the correctness of the numerals based on context. For example, to determine the numbers in “Toddler Left in Car Puts It in Drive, Hits ____ Cars”, the article mentioned “left his 2-year-old daughter alone” and “car struck two other vehicles”. Although these numerals represent different meanings, in this news headline, we knew that we should use “Trans(two)” instead of “Copy(2)” to obtain the correct result.

3.2. Dataset Construction

To derive the equations necessary for computing the correct numeral in the headline, we engage annotators via the *Amazon Mechanical Turk* platform. We formulate a question by randomly omitting one number in the headline. The annotators are then presented with the news article and corresponding question, and they must determine whether the answer is inferable from the content. If the answer is unobtainable, annotators are required to provide a detailed reason, and we designate this instance as an unanswerable question. Conversely, if the answer can be inferred, annotators utilize the predefined operators, including *Copy*, *Trans*, *Span*, *Round*, *Paraphrase*, *Add*, *Subtract*, *Multiply*, and *Divide*, to formulate an equation that yields the answer.

We enforce annotation quality via an automated validation method. Given that the ground truth is formulated by professional journalists, we need to ensure that the annotator’s equation aligns with it. When an annotator submits her/his equation, our program automatically calculates the result and checks for consistency with the article’s numerals and text spans. In essence, an annotation will be successfully submitted if its result matches the ground truth and all used numbers and text spans appear in the article. Otherwise, annotators are prompted to review their work. While this automated method effectively filters obvious errors, it is incapable of distinguishing instances where all numbers are present in the article and the equation matches the ground truth. Thus, we deploy human validation to further verify the annotations in their context. For this task, we engage 840 experienced Turkers with a hit approval rate of no less than 85% on the MTurk platform. We pay $0.45 for each annotation, and each task is randomly assigned to three different annotators. An assignment is approved if at least two annotators concur on the answer. If a consensus is not reached, the assignment is reassigned to three new annotators.

We computed the Fleiss Kappa (K) scores as a measure of inter-annotator agreement. For annotations on Amazon Mechanical Turk, we engaged three distinct annotators to validate the same task, yielding a K of 0.7753. Data annotation consistency was higher when directly copying numbers from news articles, with $K = 0.8859$. Conversely, for annotations requiring numerical reasoning, the consistency score was lower, with $K = 0.7124$.

3.3. Dataset Analysis

The proposed NumHG dataset is distinguished by three salient characteristics, as demonstrated in Table 3. Firstly, it exhibits considerably larger average sentence and word counts compared to its counterparts. Secondly, NumHG’s source articles contain more numerals than those in preceding datasets. Finally, unlike other works, NumHG incorporates unanswerable questions, with annotators asked to provide a rationale for their unanswerability. This unique feature establishes a preliminary exploration of unanswerable questions in numeral problem-solving scenarios. As depicted in Table 2, the *Copy* operator is the most commonly applied in the news articles within NumHG. The prevalence of simple operations (*Copy*, *Trans*, *Span*, *Round*, and *Paraphrase*) underscores the journalistic practice of clear information delivery, avoiding any unnecessary challenge to the reader’s numeracy skills. This also marks a notable departure from prior numerical reasoning datasets (Huang et al., 2016; Ling et al., 2017; Wang et al., 2017; Amini et al., 2019; Patel et al., 2021), which predominantly aim to assess machine numeracy, thus not directly applicable to the news article context.

3.4. Example of Unanswerable Question

In a news text like the following: “(Apr 9, 2014 4:34 PM CDT) A vehicle crashed into an Orlando-area day care today, killing one child and injuring about a dozen more, reports the Orlando Sentinel. At least one adult was reported to be injured as well. Several of the victims were in serious condition hours after this afternoon’s accident in Winter Park. Police say a Dodge Durango crashed into another vehicle, which went out of control and smashed into the KinderCare building. The driver of the Durango fled the scene in his vehicle, though police later found it in Winter Park. Authorities have named 26-year-old Robert Corchado as a person of interest.” When we tried to calculate the number underlined in “1 Child Dead, ____ Injured When Car Hits Day Care”, the original de-

	Num Acc.			ROUGE			BERTScore			MoverScore
	Overall	Copy	Reasoning	1	2	L	P	R	F1	
BART	**70.09**	**73.88**	**61.54**	46.63	21.79	41.55	48.02	49.19	48.62	62.57
T5	67.84	71.42	59.74	47.82	23.10	42.89	50.23	49.64	49.94	62.98
Pegasus	66.45	70.25	57.86	48.08	23.40	43.25	50.97	49.99	50.49	63.11
Season	67.81	71.11	60.35	48.58	23.81	43.74	51.64	50.32	50.98	63.29
BRIO	66.56	70.43	60.07	**48.93**	**24.09**	**44.12**	**52.17**	**50.84**	**51.43**	**63.50**

Table 4: Automatic evaluation results.

	Num Acc.	Reasonableness	Readability
BART	59.2	43.9	53.7
T5	53.9	52.1	55.9
Pegasus	64.6	58.8	61.2
Season	62.7	63.6	60.7
BRIO	**79.1**	**65.2**	**63.5**

Table 5: Human evaluation results.

scription in the text was "killing one child and injuring about a dozen more", without specifying the exact number of injured individuals. However, the ground truth in the headline was 10, and since we couldn't obtain the correct answer from the news text, it was deemed an unanswerable question after human validation by three annotators.

4. Experimental Evaluation

4.1. Experimental Setup

Dataset Description To facilitate equitable comparisons, we employ 5-fold cross-validation on NumHG and report the averaged results. Each fold of the NumHG dataset is partitioned into 19,422 training pairs, 2,775 validation pairs, and 5,549 test pairs.

Models We evaluate a selection of robust baseline models on our proposed dataset. Specifically, we utilize *BART* (Lewis et al., 2020), *T5* (Raffel et al., 2020), and *Pegasus* (Zhang et al., 2020a), all renowned, large-scale, pre-trained sequence-to-sequence generation models. Season (Wang et al., 2022) applies ROUGE-L between each document sentence and its corresponding reference summary to denote sentence informativeness, which subsequently guides abstractive summarization. *BRIO* (Liu et al., 2022) combines contrastive and cross-entropy losses to optimize both token-level prediction accuracy and sequence-level coordination.

Evaluation Metrics We employ ROUGE (Lin, 2004) as the automatic evaluation metric, incorporating ROUGE-1, ROUGE-2, and sentence-level ROUGE-L. We also assess baseline performance using two model-based semantic similarity metrics, BERTScore (Zhang et al., 2020b) and MoverScore (Wei Zhao, 2019). Specifically, we use *distilbert-base-uncased* to calculate MoverScore and report the Precision (P), Recall (R), and F1 measure (F1) of BERTScore through *roberta-large*.

Implementation Details In our experiments, we fine-tune *BART-large*, *T5-large*, and *Pegasus-large* from the *transformers* (Wolf et al., 2020) library. We apply beam search with a beam size of 4 and set our batch size to 16 to fully utilize GPU memory. We employ the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 5e-5. The models are trained for 15 epochs in each fold, with the performance on the validation set guiding the checkpoint selection. For Season, we use *BART-large* as a backbone, with all other settings consistent with the aforementioned. In the case of BRIO, we use *BART-large* and employ beam search to generate 16 candidate summaries. The batch size is set to 16 for a total of 30 epochs, with the Adam optimizer and learning rate scheduling following the original paper's specifications. Additionally, we employ a linear warmup strategy, setting the number of warmup steps to 2,000. All experiments were conducted on 4 NVIDIA Tesla V100 (32G) GPUs.

4.2. Experimental Results

Our principal results are displayed in Table 4. Among all the baseline models, BRIO exhibits superior performance on three summarization evaluation metrics: the ROUGE score, BERTScore, and MoverScore. Season utilizes the informativeness of each sentence to guide abstractive summarization, yielding promising improvements over the original BART by 1.95/2.02/2.19 points for ROUGE-1/2/L scores, 3.63/1.13/2.36 points for BERTScore-P/R/F1, and 0.72 points for MoverScore. Compared to Season, BRIO achieves marginal enhancements of 0.35/0.28/0.38 points for ROUGE-1/2/L scores, 0.53/0.52/0.45 points for BERTScore-P/R/F1, and 0.21 points for MoverScore. We provide not only summarization evaluation scores but also numeral accuracy (Num Acc.) to assess whether the numerals generated in the headline match those in the ground truth. *Overall* demonstrates performance on all test questions. *Copy* denotes performance on questions where answers can be directly copied from the given article. *Reasoning* pertains to questions that necessi-

Numeral Error Types	Copy(%)	Reasoning(%)
There are more than one numeral generated in the headline	13.89	9.28
There is no numeral appeared in the headline	23.55	11.84
The numeral in the headline was intended to be obtained by copying, but it was instead inferred.	24.93	/
The numeral in the headline was intended to be inferred, but it was obtained by copying.	/	**46.24**
The numeral in the headline is copied from other numbers in the article	**37.63**	/
The numeral in the headline is inferred from other numerals in the article or calculated incorrectly	/	32.64

Table 6: Error analysis in automatic evaluation.

tate numerical reasoning to derive the answer. Although BRIO excels in three summarization evaluation metrics, Table 4 reveals that BART is the most effective in generating accurate numerals in the headline, with an overall numeral accuracy of 70.09%.

4.3. Human Evaluation

We engaged five graduate students in communication and media studies as annotators, randomly selecting 100 instances from the NumHG test set. Each annotator was provided with the news article and five generated headlines, with no knowledge of which model generated which headline. We requested them to evaluate the generated headlines on three criteria: *Numeral Accuracy*, *Reasonableness*, and *Readability*. Numeral Accuracy assesses the correctness of numbers in the headline. The score is categorized as follows: 0 indicates all numbers in the generated headline are incorrect, 1 indicates a portion of numbers are correctly predicted, and 2 indicates all numbers are correctly predicted. Reasonableness measures whether the generated headlines are suitable for the news context and requires the annotators to select the best headline for the given article. The best headline score 5, the second-best score 4, and the least favored score 1. Readability measures the ease or difficulty of understanding the headline. The readability score ranges between 1 and 5, where 1 signifies the generated headline is very challenging to read, and 5 indicates the headline is easily readable and understandable. Lastly, we represent the human evaluation results as percentages. For instance, we first sum up the numeral accuracy score of BART given by each evaluator. Then, we obtain 118.4, which is the average numeral accuracy score of the five annotators. Finally, we convert the numeral accuracy score into a percentage as 118.4/200, where 200 is the maximum possible score for the 100 sampled instances. The calculated Kappa scores for these aspects were $K_{\text{accuracy}} = 0.8527$, $K_{\text{reasonableness}} = 0.7297$, and $K_{\text{readability}} = 0.7809$.

Table 5 reports the human evaluation results. As illustrated in Table 5, BRIO excels in Numeral Accuracy, Reasonableness, and Readability in human evaluations. Interestingly, the numeral generated in the headline by baseline models can also be correct, even if it does not match the ground truth. However, a large number of numerals in generated headlines could be incorrect if the numeral's context is taken into account.

4.4. Error Analysis

We manually inspected the cases with incorrect numerals generated in the headlines by BRIO model. Table 6 shows the results of error analysis in automatic evaluation. As illustrated in Table 6, for those cases where numerals need to be directly copied from the given article, 37.63% errors occur when the numeral in the headline is copied from other numbers in the article. Besides, nearly 25% errors occur if the numeral in the headline was intended to be obtained by copying, but it was instead inferred. As for those cases that require numerical reasoning to derive the numerals in the headlines, 46.24% of the errors are due to lack of numerical reasoning and instead directly copying numbers from the given article.

5. Conclusion

This paper concentrates on numerals when generating headlines and introduces a challenging dataset, NumHG. We employ several state-of-the-art models to generate headlines containing accurate numerals. However, experimental results indicate that these robust baseline models fail to generate accurate headlines with correct numerals. We will release NumHG under the CC BY-SA 4.0 license.[2]

6. Acknowledgement

This work was supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3, NSTC 112-2634-F-002-005 -, and Ministry of Education (MOE) in Taiwan, under grants NTU-112L900901. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

[2] NumHG: https://github.com/ArrowHuang/NumHG

7. References

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2357–2367.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Nquad: 70,000+ questions for machine comprehension of the numerals in text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2925–2929.

Jui Chu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Learning to generate correct numeric values in news headlines. In *Companion Proceedings of the Web Conference 2020*, pages 17–18.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4830–4842.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1074–1084. Association for Computational Linguistics.

Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 887–896.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

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.

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

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 158–167.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2890–2903.

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.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2080–2094.

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

Fei Wang, Kaiqiang Song, Hongming Zhang, Lifeng Jin, Sangwoo Cho, Wenlin Yao, Xiaoyang Wang, Muhao Chen, and Dong Yu. 2022. Salience allocation as guidance for abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6094–6106.

Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 845–854.

Fei Liu Yang Gao Christian M. Meyer Steffen Eger Wei Zhao, Maxime Peyrard. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 563–578.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 38–45.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

8. Appendix

We provide some case studies in this section. Here are examples where the generated headline contains the correct numerals but does not match the ground truth:

News: (Jun 21, 2013 9:57 AM CDT) Yesterday saw the completion of the 90th National Marbles Tournament, and America has crowned two new winners. Emily Cavacini, 11, who hails from outside Pittsburgh, took home the title of girls' champion, while Cooper Fisher, 12, from Middletown Valley, Maryland, beat out the other boys, reports the AP. It was a more involved process than you might imagine: Each competed against 25 others in a tournament in Wildwood, NJ, that lasted four days. The Post-Gazette reports that more than 1,200 (yes, twelve-hundred) games were played, in which the kids were tasked with knocking seven marbles out of a circle using a shooter marble.
Ground Truth: 1,200 Games Later, America Has New Marbles Champs
Generated Headline 1: America's 90th National Marbles Tournament Is Over
Generated Headline 2: America Crowns 2 New Marbles Champs

Table 7: An example of correctly generated numerals that do not match the ground truth.

As shown in Table 7, in generating the first headline, the model focused on which edition of the National Marbles Tournament it was. In generating the second headline, the model emphasized the emergence of champions. Although the numerals in both generated headlines are correct, they do not match the ground truth.

Here are examples where the generated headline is incorrect due to contextual considerations:

News: (Jan 11, 2011 7:01 PM) Here's one more reason California finds itself with a $25 billion budget deficit—it was handing out cell phones to state employees like candy. Nearly 100,000 state workers—about 40% of the workforce—have phones on the taxpayers' dime. New Gov. Jerry Brown is ordering that half of them be turned in, for an estimated savings of about $20 million a year, reports the Los Angeles Times. 'It is difficult for me to believe that 40% of all state employees must be equipped with taxpayer-funded cellphones,' said the governor. Some state employees, including department and agency executives who are required to be in touch 24 hours a day and seven days a week, may need cellphones, but the current number of phones out there is astounding.
Generated Headline 1: 40% of Californians Have Taxpayer-Funded Cellphones
Generated Headline 2: State Workers Get Cell Phones on Taxpayer-Funded $25B

Table 8: An example of incorrectly generated numerals if the context is taken into account.

As illustrated in Table 8, taking into account the entire news text, the first headline is incorrect; it should be "40% of Californian state employees have taxpayer-funded cellphones." The second headline is also incorrect because the news text mentions "one more reason of $25 billion budget deficit was handing out cell phones to state employees like candy." In summary, combining the headline with the context can help determine that the generated numerals are incorrect.