Sequential Text-based Knowledge Update with Self-Supervised Learning for Generative Language Models

Hao-Ru Sung
National Chengchi University
Taipei, Taiwan
110753124@nccu.edu.tw

Ying-Jhe Tang
Academia Sinica
Taipei, Taiwan
isaac0511@iis.sinica.edu.tw

Yu-Chung Cheng
National Chengchi University
Taipei, Taiwan
yuchungc@nccu.edu.tw

Pai-Lin Chen
National Chengchi University
Taipei, Taiwan
pailinch@nccu.edu.tw

Tsai-Yen Li
National Chengchi University
Taipei, Taiwan
li@cs.nccu.edu.tw

Hen-Hsen Huang
Academia Sinica
Taipei, Taiwan
hhhuang@iis.sinica.edu.tw

ABSTRACT

This work proposes a new natural language processing (NLP) task to tackle the issue of multi-round, sequential text-based knowledge update. The study introduces a hybrid learning architecture and a novel self-supervised training strategy to enable generative language models to consolidate knowledge in the same way as humans. A dataset was also created for evaluation and results showed the effectiveness of our methodology. Experimental results confirm the superiority of the proposed approach over existing models and large language models (LLMs). The proposed task and model framework have the potential to significantly improve the automation of knowledge organization, making text-based knowledge an increasingly crucial resource for powerful LLMs to perform various tasks for humans.

CCS CONCEPTS

• **Information systems** → *Digital libraries and archives*; • **Computing methodologies** → **Language resources**; **Natural language generation**; **Discourse, dialogue and pragmatics**.

KEYWORDS

natural language generation, temporal knowledge modeling, update summarization, self-supervision

ACM Reference Format:
Hao-Ru Sung, Ying-Jhe Tang, Yu-Chung Cheng, Pai-Lin Chen, Tsai-Yen Li, and Hen-Hsen Huang. 2023. Sequential Text-based Knowledge Update with Self-Supervised Learning for Generative Language Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615188

1 INTRODUCTION

The rapid pace of change in the world means that knowledge is constantly evolving, and that existing knowledge carriers need to be updated in order to stay relevant and accurate. As pointed out by previous work [15], keeping knowledge up-to-date is a complex and time-consuming task, particularly for large knowledge carriers like Wikipedia, which is maintained by a large team of editors working to keep its 6.5 million articles up-to-date.

This work addresses the task of multi-round sequential knowledge update in a time window. This is a special case of knowledge update frequently occurs in the real world. For example, when a flood just occurs, there may be initial, real-time reports, but as time goes by, the disaster will continue to spread, and as more journalists get involved, more details about the event will be revealed. During the days following the flood, the knowledge carrier must continue to receive updates on the disaster situation, supplementing more and more details, and even correcting previous reporting errors.

In addressing this issue, we investigate a new natural language processing (NLP) task that extracts relevant information from continuous, constantly updating news sources to keep knowledge up-to-date. In current information systems, knowledge was stored in different ways. Besides text, there were symbolic, graph-like knowledge bases or numerical, vector-like knowledge representations. These structured data were easier to process than unstructured, discrete text information. However, in the process of being transformed into knowledge bases or knowledge representations, information loss was inevitable. Additionally, it was also difficult for humans to read these transformed contents.

With the rise of large language models (LLMs) such as T5 [21] and GPT [5], these text-to-text models have natural language as the medium for input and output. As a result, storing knowledge directly in natural language form allows computers to perform understanding and inference intelligently, while avoiding the limitations of transformation distortion and human difficulty in reading. Therefore, in this study, we use natural language as the representation of knowledge and perform knowledge update tasks.

We propose a hybrid learning architecture and a novel self-supervised training strategy for an LLM to capture the way humans consolidate knowledge. To the best of our knowledge, this is the first study on the task of text-based sequential knowledge update. For this reason, we also develop a dataset for evaluation. Experimental results show the effectiveness of our methodology. We also evaluate the impressive LLM-based service, ChatGPT and analyzed the outputs from various models to confirm the superiority of our approach over the powerful ChatGPT model in this very task.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615188

Overall, our proposed task and model framework have the potential to significantly enhance the automation of knowledge organization. As text-based knowledge is becoming an increasingly crucial resource for powerful LLMs to perform various tasks for humans, it is essential to ensure the availability and accuracy of real-time updated knowledge. This pilot study demonstrates a promising research direction for addressing these challenges. The contributions of this work are threefold as follows:

- Introduction of a new NLP task: The study introduces a new NLP task to tackle the issue of sequential text-based knowledge update, which is a critical research task in the era of widespread use of LLMs in high-intelligence tasks.
- Proposal of a hybrid learning architecture and a novel self-supervised training strategy: A hybrid learning architecture and a novel self-supervised training strategy are proposed to capture the way humans consolidate knowledge. Experimental results confirm the effectiveness and superiority of the approach over existing models and the LLM ChatGPT.
- Development of a dataset: A dataset was created for this pilot task. The dataset is challenging for ChatGPT and can serve as a valuable research resource for the community.[1]

2 RELATED WORK

Prior research has explored the temporal dynamics of knowledge through methods like constructing knowledge bases [23, 24], addressing the update summarization task [1, 2, 8, 18], and summarizing news articles [9]. Notably, Dang et al. [8] introduced the update summarization task to create follow-up news summaries from a collection of articles. Some studies have also tackled news headline generation, using a news article as input [4, 10, 17, 22, 26].

Our approach contrasts with update summarization in its objectives and formulation. Unlike update summarization, which emphasizes contrasting new information for follow-up, our task seeks to blend both the new and existing data in an updated text. Moreover, while update summarization uses a multi-document foundation based on prior summaries, we aim to organize sequential updates for the same target knowledge. Our endeavor also varies from updated headline generation [18], as we handle longer texts and ensure a balanced incorporation of both old and new information, contrasting the latter's emphasis on differences.

Compared to prior research on news event-triggered knowledge updates that focus on single-round updates [15], our approach delves into multiple sequential knowledge updates, proposing a novel method that exhibits significant potential for the task.

3 PROBLEM FORMULATION

The task addressed in this work can be specified as follows. Given $\langle x_1, x_2, ..., x_{t-1} \rangle$, a sequence of text-based knowledge facts until $t-1$ and e_t, a piece of information about a related event that occurs at time t, we aim to generate x_t, an updated version of x_{t-1} with the acknowledgement of e_t. According to the statistics in Table 1, each cluster has a mean of 60 news events, so the average length of our sequence is the same as 60. For this reason, a text-to-text model for this task can be formulated as Equation 1 and approximated by a conditional generation model as Equation 2.

$$x_t = \arg\max_y Pr(y|x_1, x_2, ..., x_{t-1}, e_t) \quad (1)$$

$$\approx \arg\max_y Pr(y|x_{t-1}, e_t) \quad (2)$$

Note that our task is aimed at generating the entire sequence of $x_2, x_3, ..., x_t$ iteratively, given the initial version x_1. Thus, this task suffers from the issue of error propagation. That is, the incorrect generation in x_i may cause the error in x_{i+1}. To investigate this task, we create a dataset for experiments as described in Section 4 and propose a hybrid learning approach with self-supervision as introduced in Section 5.

4 DATASET CONSTRUCTION

In this work, we regard the summary of a Wikipedia page as a knowledge fact. The different editing versions of a Wikipedia page are taken as the sequential data $\mathbf{x} = \langle x_1, x2, ..., x_T \rangle$ of T updates. To align $\mathbf{x}$ with the sequence of event information $\mathbf{e} = \langle e_1, e_2, ..., e_T \rangle$, we exploit the update history of the Wiki Current Event Portal,[2] and fetch the cited news articles from Common Crawl[3] that contain the information of $\langle e_1, e_2, ..., e_T \rangle$. The details are described as follows.

Inspired by the concept in Gholipour Ghalandari et al. [11], we first collect the news events that are listed in Wiki Current Event Portal and Common Crawl, then we collect the Wikipedia page which is a topic that related to the news articles. For each knowledge fact, we can obtain a cluster of news articles $e_1, e_2, e_3, ..., e_T$ that are related to the knowledge fact, and the timestamps $1, 2, ..., T$ of news articles published. We find the pair of articles (x_t, x_{t+1}) from Wikipedia, which x_t is the version that edited before t , and x_{t+1} is the version that edited after t.

Our dataset consists of numerous knowledge facts. Each knowledge fact c has multiple versions of news articles $\langle e_1, e_2, e_3, ..., e_T \rangle$ from Wiki Current Event Portal and Common Crawl, and a serious of timestamps $1, 2, ..., T$ which are correspond to the published time of news articles. For a news article in c which comes from Wiki Current Event Portal, we collect the Wikipedia page linked from the portal of Wiki Current Event. The Wikipedia page may have different versions, we find the latest version that the edit time is prior to t which is consider as original Wikipedia page before the occurrence of e_t. And we find the first Wikipedia page that the edit time is after t which is consider as an updated version due to the multiple versions of news articles. The resulting pair of sequences $(\mathbf{x}, \mathbf{e})$ is considered as an instance of our dataset.

We have retained the data that the text was written in English. The constructed dataset contains a total of 5,695 instances. Based on the chronological order, the first 80% of the instances are assigned to the training set, the next 10% are designated as the validation set, and the remaining 10% are used for testing. Table 1 shows the statistics of our dataset.

5 METHODOLOGY

Our framework consists of two steps: teacher training and student training. We refer to this framework as a hybrid training architecture, in which teacher training is considered the pre-training stage,

[1] https://github.com/HaoruSung/Sequential-Text-based-Knowledge-Update.git

[2] https://en.wikipedia.org/wiki/Portal:Current_events

[3] https://commoncrawl.org/2016/10/news-dataset-available/

	Training	Validation	Test
# Clusters	4,583	556	556
# Articles	276k	43k	42k
Mean	60.3	76.8	74.9
Median	66	100	100
Period Begin	2016-8-25	2019-1-6	2019-5-8
Period End	2019-1-5	2019-5-7	2019-8-20

Table 1: Statistics of our dataset

Dataset	# Inst.	Avg. Sent.	Avg. Tokens
Labeled data (NetKu)	253	12.27	244.60
Unlabeled data (WCEP)	1,990	9.47	193.11
Total	2,243		

Table 2: Statistics of the data for teacher training

and student training belongs to the fine-tuning stage. The backbone model is T5-Base [21].

5.1 Hybrid Training Framework

Figure 1 illustrates the hybrid training architecture. During teacher training, the model uses unlabeled data for self-supervised training under different pre-training strategies. For student training, the model is first initialized with pre-training parameters, then all parameters are fine-tuned using labeled data from the knowledge update task.

Student Training Stage. This stage fine-tunes the model using our dataset. Assuming that event e is dynamic and occurs in a temporal order. When an event occurs, new information, ranging from e_1 to e_t, emerges over time. As input, the e_t data at timestamp t is paired with the x_t information that existed prior to the event. The intended output is x_{t+1}, inclusive of the new e_t information. The subsequent timestamp e_{t+1} prompts an update to the x_{t+1} text. The output from the previous timestamp serves as the input for the next timestamp, until the final version is generated. In this way, the model is trained to iteratively update text knowledge in scenarios where new information continuously emerges. When the model is trained exclusively through the student training stage, we observe that the model tends to forget previous information as t increases. Therefore, we implement teacher training to overcome the limitations encountered by the model.

Teacher Training Stage. Teacher training is further divided into two steps. First, we use the NetKu [15] and WCEP [11] datasets that reappear in the task validation set and test set. The remaining data are used as training data for the teacher training stage. However, training the model with only a small amount of data the NetKu dataset is insufficient to comprehensively capture new knowledge. Due to the lack of relevant large-scale datasets, we additionally crawled more entries from WCEP to augment the training data as the material for self-supervised pre-training. The statistics of the pre-training data are given in Table 2.

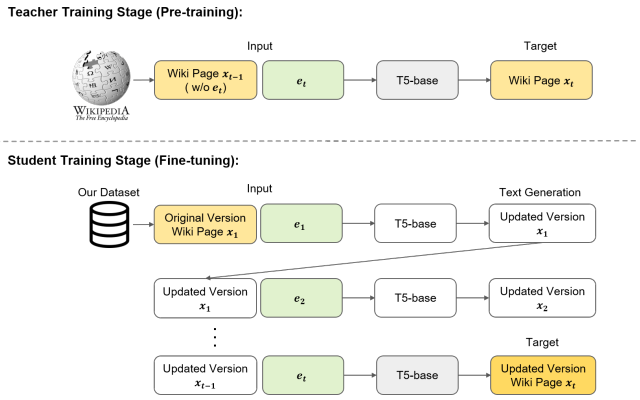

Figure 1: Hybrid Training Architecture.

5.2 Self-Supervised Pre-training

While facing data scarcity, the task of textual knowledge updating is similar to reconstructing text from a corrupted one in a way. Specifically, the first paragraph of a Wikipedia page is selected as x_t. From x_t, 15% of the sentences are randomly extracted as the new information e_t, and any sentences in x_t that duplicate with e_t are subsequently removed to form x_{t-1}. The concatenation of x_{t-1} and e_t into an input through prompts in the format of "update: $\{x_{t-1}\}$ event: $\{e_t\}$", using x_t as reference output. Consequently, it is possible to obtain the triplet of (x_{t-1}, e_t, x_t) as candidate instances.

We explore various perturbation strategies to pre-train our model in text reconstruction, as summarized in Table 3.

(1) **Sentence-Shuffle (SS)**: e_t is a part of the sentences from x_t and the sentence order is shuffled.
(2) **Noise-Injection (NI)**: it is based on Random-Event, but additional noise is introduced into e_t, where the noise comes from the content of other Wikipedia pages irrelevant to x_t, and the added noise accounts for 20% of e_t.
(3) **Masked-Noise-Injection (MNI)**, based on NI, we implement an extreme and selective masking strategy following [12]. In the extraction step, we use the unsupervised keywords extraction algorithm YAKE [6] to extract keywords or key-phrases (up to 3-grams) from the x_{t-1}, constituting approximately 20% of the entire text length. The masking ratio is 15%, consistent with that adopted in the original T5 pre-training.
(4) **Noise-Generation (NG)**, it is similar to NI, but the noise is generated by GENIUS [12]. The GENIUS model generates text related to x_t based on the original text, and we extract 20% of the sentences from it as noise.
(5) **Masked-Noise-Generation (MNG)**, based on NG, some tokens are further masked by the strategy described in MNI.

6 EXPERIMENTS

We evaluate our hybrid framework with a variety of self-supervised pre-training strategies. The metrics include ROUGE [16], BLEU [19], METEOR [3], and BERTScore [27]. Our approach is compared with a student model in the typical supervised-learning. We further

Perturbation Strategy	Scenario in text reconstruction
Original (No Perturbation)	To update x_{t-1} (Wiki Content) with the event e_t (original news article)
Sentence-Shuffle (SS)	To update x_{t-1} with event e_t that is randomly shuffled
Noise-Injection (NI)	To update x_{t-1} with event e_t that is injected with noise from irrelevant Wiki content
Masked-Noise-Injection (MNI)	To update masked x_{t-1} with event e_t that is injected with noise from irrelevant Wiki content
Noise-Generation (NG)	To update x_{t-1} with event e_t that is augmented with noise generation
Masked-Noise-Generation (MNG)	To update masked x_{t-1} with event e_t that is augmented with noise generation

Table 3: Perturbation strategies to create self-supervised data for pre-training language models in text reconstruction

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	BERTScore
Student Mode	15.14	7.61	12.26	3.67	2.31	1.98	1.83	7.50	57.68
Hybrid (Original)	67.94	63.50	65.63	52.22	49.65	48.46	47.53	57.53	84.95
Hybrid (Original+SS)	74.02	70.46	72.05	63.22	61.05	59.82	58.85	66.93	87.85
Hybrid (Original+NI)	85.19	81.81	83.23	77.15	74.69	73.25	72.13	79.76	92.94
Hybrid (Original+MNI)	78.61	74.63	76.56	68.52	66.06	64.66	63.56	71.21	90.04
Hybrid (Original+NG)	83.32	79.86	81.39	74.97	72.50	71.09	69.95	77.71	92.03
Hybrid (Original+MNG)	84.45	80.88	82.38	76.64	74.07	72.73	71.70	79.31	92.62
Hybrid (Original/NI+MNG)	**86.44**	**83.02**	**84.46**	**79.63**	**77.08**	**75.65**	**74.54**	**82.16**	**93.49**
GPT-3.5-turbo (Zero-shot)	49.35	34.98	40.41	39.79	31.71	30.07	29.59	41.00	72.75
Falcon-7B (Zero-shot)	15.96	2.08	10.14	8.16	0.99	0.13	0.02	8.36	42.42
LLaMA-13B (Zero-shot)	16.60	9.28	13.40	9.53	7.57	7.31	7.26	10.77	47.25
Vicuna-13B (Zero-shot)	43.94	28.68	35.67	36.18	26.75	25.17	24.89	35.97	68.16
Falcon-7B (Fine-Tuned)	21.20	4.66	12.32	10.79	2.19	0.53	0.18	11.13	55.38
LLaMA-13B (Fine-Tuned)	32.53	22.55	27.66	24.01	17.72	16.32	15.66	30.19	60.00
Vicuna-13B (Fine-Tuned)	48.00	36.25	41.76	42.00	34.28	33.22	33.04	42.55	68.98

Table 4: Experimental results. The "+" symbol represents a mixture of data from multiple strategies, while the "/" symbol represents a phased training of data from multiple strategies.

compared our approach with four LLMs: GPT-3.5-turbo, LLaMA-13B [25], Vicuna-13B [7], and Falcon-7B [20]. We explored a zero-shot prompting [14] method, allowing the four LLMs to directly infer, given x_{t-1} and e_t to generate x_t. Moreover, we use LoRA [13] to fine-tune the three LLMs with our top-performing strategy combination (Original/NI+MNG) followed by multi-round inference.

6.1 Results

Experimental results are given in Table 4. All the hybrid approaches surpass the traditional supervised-learning student model. In addition, our pre-training strategies further improve the backbone generative language model for generating updated knowledge facts. The masking strategy is not always effective. That is, MNG is better than NG, while MNI is inferior to NI. Thus, we further combine MNG and NI for self-supervised learning, and the experimental result confirms that the hybrid model (Original/NI+MNG) achieves the best performance in all metrics. The results suggests that both kinds of noise information successfully improve the T5 model's ability to handle updated information.

Our premier model consistently outperforms all the four LLMs across all evaluation metrics. The models associated with these four LLMs frequently produce results that are lengthy and redundant. In contrast, our model delivers outputs that are notably more precise and topical. In the Zero-shot prompting experiments, while GPT-3.5-turbo exhibits commendable performance, it still does not match the efficiency of our leading model. After the fine-tuning through LoRA, we observed the improvements across all LLMs. This progression further underscores the efficacy of our training strategy.

7 CONCLUSIONS

This work addresses the challenge of sequential text-based knowledge update in a time window. The aim is to extract relevant information from constantly updating news sources to keep the knowledge updated. The study proposes a hybrid learning architecture and a novel self-supervised training strategy for LLMs to consolidate knowledge in a way similar to humans. The proposed task and model framework have the potential to significantly improve the automation of knowledge organization.

ACKNOWLEDGMENTS

This work is partially supported by National Science and Technology Council, Taiwan under grants 109-2222-E-001-004-MY3 and 112-2221-E-001-016-MY3 and by Academia Sinica under grants 3006-37C4527 and 3006-37C4386.

REFERENCES

[1] Javed Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Tetsuya Sakai. 2015. *Trec 2014 temporal summarization track overview*. Technical Report. NATIONAL INST OF STANDARDS AND TECHNOLOGY GAITHERSBURG MD.

[2] Javed A Aslam, Matthew Ekstrand-Abueg, Virgil Pavlu, Fernando Diaz, and Tetsuya Sakai. 2013. TREC 2013 Temporal Summarization.. In *TREC*.

[3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[4] Michele Banko, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline Generation Based on Statistical Translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (Hong Kong) *(ACL '00)*. Association for Computational Linguistics, USA, 318–325. https://doi.org/10.3115/1075218.1075259

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword Extraction from Single Documents Using Multiple Local Features. *Inf. Sci.* 509, C (jan 2020), 257–289. https://doi.org/10.1016/j.ins.2019.09.013

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[8] Hoa Trang Dang, Karolina Owczarzak, et al. 2008. Overview of the TAC 2008 update summarization task.. In *TAC*.

[9] Franck Dernoncourt, Mohammad Ghassemi, and Walter Chang. 2018. A Repository of Corpora for Summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1509

[10] Bonnie J Dorr, David Zajic, and Richard Schwartz. 2003. Hedge Trimmer: a parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop - Volume 5 (HLT-NAACL-DUC '03)*. Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 1 – 8. https://doi.org/10.3115/1119467.1119468

[11] Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A Large-Scale Multi-Document Summarization Dataset from the Wikipedia Current Events Portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1302–1308. https://doi.org/10.18653/v1/2020.acl-main.120

[12] Biyang Guo, Yeyun Gong, Yelong Shen, Songqiao Han, Hailiang Huang, Nan Duan, and Weizhu Chen. 2022. GENIUS: Sketch-based Language Model Pre-training via Extreme and Selective Masking for Text Generation and Augmentation. arXiv.

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL]

[14] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916 [cs.CL]

[15] Yu-Ting Lee, Ying-Jhe Tang, Yu-Chung Cheng, Pai-Lin Chen, Tsai-Yen Li, and Hen-Hsen Huang. 2022. A Multi-Grained Dataset for News Event Triggered Knowledge Update. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 4158–4162. https://doi.org/10.1145/3511808.3557537

[16] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[17] 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*. Association for Computational Linguistics, Online, 1335–1346. https://doi.org/10.18653/v1/2020.acl-main.123

[18] Sheena Panthaplackel, Adrian Benton, and Mark Dredze. 2022. Updated Headline Generation: Creating Updated Summaries for Evolving News Stories. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 6438–6461. https://aclanthology.org/2022.acl-long.446

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[20] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv:2306.01116 [cs.CL]

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 1, Article 140, 67 pages.

[22] Sho Takase, Jun Suzuki, Naoaki Okazaki, Tsutomu Hirao, and Masaaki Nagata. 2016. Neural Headline Generation on Abstract Meaning Representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1054–1059. https://doi.org/10.18653/v1/D16-1112

[23] Xavier Tannier and Véronique Moriceau. 2013. Building Event Threads out of Multiple News Articles. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 958–967. https://aclanthology.org/D13-1098

[24] Sansiri Tarnpradab, Fereshteh Jafariakinabad, and Kien A Hua. 2021. Improving online forums summarization via hierarchical unified deep neural network. *arXiv preprint arXiv:2103.13587* (2021).

[25] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[26] David Zajic, Bonnie Dorr, and Richard Schwartz. 2002. Automatic Headline Generation for Newspaper Stories. In *IN THE PROCEEDINGS OF THE ACL WORKSHOP ON AUTOMATIC SUMMARIZATION/DOCUMENT UNDERSTANDING CONFERENCE (DUC*. 78–85.

[27] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).