Time-aware Prompting for Text Generation

Shuyang Cao and **Lu Wang**
Computer Science and Engineering
University of Michigan
Ann Arbor, MI
{caoshuy, wangluxy}@umich.edu

Abstract

In this paper, we study the effects of incorporating timestamps, such as document creation dates, into generation systems. Two types of time-aware prompts are investigated: (1) **textual prompts** that encode document timestamps in natural language sentences; and (2) **linear prompts** that convert timestamps into continuous vectors. To explore extrapolation to future data points, we further introduce a new data-to-text generation dataset, TEMPWIKIBIO, containing more than 4 millions of chronologically ordered revisions of biographical articles from English Wikipedia, each paired with structured personal profiles. Through data-to-text generation on TEMPWIKIBIO, text-to-text generation on the content transfer dataset, and summarization on XSum, we show that linear prompts on encoder and textual prompts improve the generation quality on all datasets. Despite having less performance drop when testing on data drawn from a later time, linear prompts focus more on non-temporal information and are less sensitive to the given timestamps, according to human evaluations and sensitivity analyses. Meanwhile, textual prompts establish the association between the given timestamps and the output dates, yielding more factual temporal information in the output.

1 Introduction

Temporal information, such as publication and modification dates of documents, is an inherent attribute of documents. Both document writers and readers are aware of this information when organizing and consuming document content. For example, an event reported by a news article is likely to happen right on the publication date. However, state-of-the-art generation models are fine-tuned from large pre-trained models without incorporating temporal information (Lewis et al., 2020; Zhang et al., 2020), creating a gap between document processing by humans and automatic models. Though previous work has split datasets according to temporal information and shown deteriorated performance of large pre-trained models as knowledge becomes outdated on future data (Lazaridou et al., 2021; Jang et al., 2022), it is unclear how informing models of temporal information affects generation tasks.

Therefore, this work aims to study the effects of presenting temporal information to generation models. Concretely, to include timestamps in model inputs, we consider prepending two types of **time-aware prompts** to the encoder or the decoder. First, **textual prompts** encode timestamps within natural language descriptions, as commonly used by the recent prompt engineering work (Radford et al., 2019; Raffel et al., 2019). We further explore **linear prompts** that map timestamps to continuous vectors via linear projections.

Concretely, we fine-tune BART (Lewis et al., 2020) with time-aware prompts and conduct experiments on two text-to-text generation tasks: (1) content transfer (Prabhumoye et al., 2019) that generates the continuation of a passage using information from a given document, and (2) summarizing news articles with XSum (Narayan et al., 2018). To study time-aware prompts' capability of extrapolating to future dates, we introduce **TEMPWIKIBIO**, a data-to-text dataset containing timestamped revisions of biographical articles from English Wikipedia, each paired with an infobox as input. The revisions record changes of personal profiles from 2004 to 2021.[1] For all experimented datasets, dated events are critical. We first evaluate model outputs with automatic metrics to examine the effects of temporal information on output informativeness. Human judges are then asked to additionally rate the factuality of model outputs and determine if the improvement or degradation is due to date changes

[1]Our data and code are available at https://shuyangcao.github.io/projects/temporal_prompt_generation.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 7231–7246
December 7-11, 2022

Infobox Attributes: name[J. Melville Broughton Jr.] term_start[December 31, 1948] death_date[March 6, 1949] ...

BART: Joseph Melville Broughton Jr. (November 17, 1888 – March 6, 1949) was the 60th Governor of the U.S. state of North Carolina from 1941 to 1945.
Linear Prompt: Joseph Melville Broughton Jr. (November 17, 1888 – March 6, 1949) was the 60th Governor of the U.S. state of North Carolina from 1941 to 1945 and a United States Senator from 1948 until his death in 1949.

BBC News: The group made a loss of $219m (£175.1m) compared with the same time last year ... This segment posted another very strong quarter ...

Original Publication Date: 2017-02-09
BART: News Corp has reported a loss for the first three months of the year.
Linear Prompt: News Corp has reported a loss for the first three months of the year.
Textual Prompt: News Corp has reported a loss for the three months to December.

Date Perturbation: 6 months after ↦ 2017-08-09
Textual Prompt: News Corp has reported a loss for the second quarter.

Date Perturbation: 1 month after ↦ 2017-03-09
Textual Prompt: News Corp has reported a loss for the first three months of the year.

Figure 1: Sample system outputs on TEMPWIKIBIO for data-to-text generation and on XSum for summarization. We highlight relevant temporal information in the input and corresponding correct (incorrect) information in the model outputs. Linear prompts could encourage selecting important dates on TEMPWIKIBIO, but the temporal information encoded in the linear prompts can not be captured by the model, leading to incorrect dates when resolving with the provided dates is required on XSum; while the model with textual prompts is sensitive to the provided dates and generates correct date, it lacks world knowledge (e.g., seasonal earning is only reported after the season) to handle the last case after perturbing the original publication date.

in the outputs. Finally, we perform a sensitivity analysis by making perturbations to the original dates (e.g., setting the dates to one year before) and providing models with the perturbed dates, and then inspect the changes of outputs. We find that:

- Time-aware prompts improve the model performance over the no-prompt baseline in 87.5% of comparisons on different metrics and datasets. Linear prompts work better on the data-to-text dataset, while textual prompts work better on the text-to-text datasets, partly due to modal compatibility.

- The improvement in output informativeness and factuality by linear prompts is less frequently related to modifying temporal information in the outputs than textual prompts, according to human judges. Moreover, models with linear prompts are less sensitive to the given temporal information, suggesting that linear prompts assist the processing of non-temporal content.

- Textual prompts associate the provided dates with the dates to be generated in the outputs, producing more factual time-related information. However, models with textual prompts could generate incorrect dates when complicated world knowledge is required to perform reasoning, as shown in the last example in Figure 1.

2 Related Work

Temporal Generalization in NLP. Early work on temporal generalization focuses on detecting the shifts of n-gram frequencies over time (Michel et al., 2011) and detecting word meaning changes (Wijaya and Yeniterzi, 2011; Kulkarni et al., 2015). Besides linguistic shifts, model degradation on downstream tasks has been reported when tested on samples at a different time from the training data (Huang and Paul, 2018; Lukes and Søgaard, 2018; Lazaridou et al., 2021; Agarwal and Nenkova, 2021). In this work, we study the temporal generalization of our time-aware prompts, since they are constructed with temporal information.

Prompt Engineering. Prompts have been a common tool for controllable generation (Fan et al., 2018; Radford et al., 2019; Keskar et al., 2019; Raffel et al., 2019). Instructions are also constructed as prompts to allow large models to perform new tasks that are unseen in training (Brown et al., 2020; Sanh et al., 2022). More recently, prompts, either hand-crafted (Schick and Schütze, 2021; Gao et al., 2021) or learned (Li and Liang, 2021), are found to benefit model learning and improve few-shot performance on downstream tasks. Our textual time-aware prompts extend the year-level prompts in Dhingra et al. (2021) with months and days to incorporate fine-grained temporal information, and we further explore representing timestamps with linear prompts which have been mainly used for length-controlled generation (Kikuchi et al., 2016).

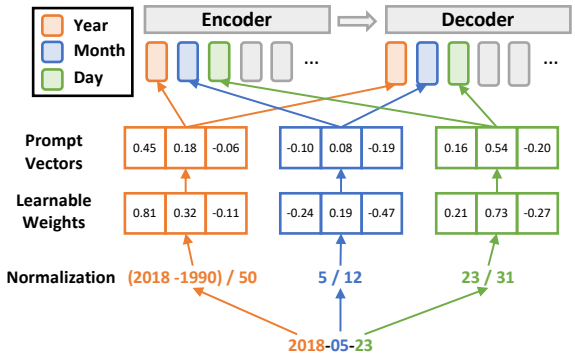

Figure 2: A linear prompt treats year/month/day as separate scalars and projects them into continuous prompt vectors to be used on encoder or decoder. The vector's scale reflects their temporal orderings. Note that the dimension of prompt vectors is the same as the embedding dimension in the actual model.

3 Time-aware Prompts

We study two types of prompts that are prepended to the encoder/decoder of a seq2seq model, to inform the model of temporal information.

Textual Prompt. Given a document's timestamp, we first convert it to "day month year" with the day and the year in digits and the month in its textual form (e.g., "*18 January 2015*"), a format commonly used by mainstream media such as BBC news. We test three textual prompts and use the one that results in the highest ROUGE score on the development set of XSum, i.e., "*Today is [timestamp].*" ("*Today is 18 January 2015.*") Other textual prompts are detailed in Appendix A. Compared to only inserting the year information (Dhingra et al., 2021), textual prompts in our paper provide more fine-grained temporal information.

Linear Prompts treat the concept of time as an axis, with each timestamp being mapped to a point on it. Concretely, we use the year, month, and day as scalars and transform them into prompt vectors through linear projections, as illustrated in Figure 2. The process of linear projections can also be viewed as changing the scales of vectors for the year, month, and day. While prior work has controlled the output lengths by changing the scales of memory cells in an LSTM (Kikuchi et al., 2016), representing temporal information with scales of vectors has yet been studied.

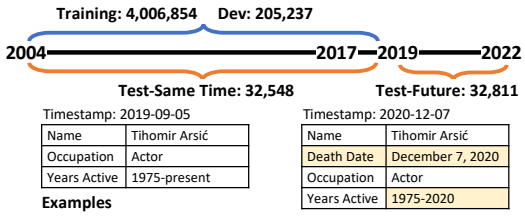

Figure 3: Numbers of samples and the corresponding time periods of revisions in different splits of TEMPWIKIBIO. Changed attributes of the sample revisions are shaded in yellow. There is **no** overlapping subject between training/development sets and the two test sets.

4 TEMPWIKIBIO Data Collection

To study how well time-aware prompts can extrapolate to future data, we collect TEMPWIKIBIO, which has 4,277,450 revisions of infobox-paragraph pairs from 2004 to 2021 for 695,929 Wikipedia biography articles, extending WIKIBIO (Lebret et al., 2016), which only includes the latest revision per article by 2015. Importantly, the profile (e.g., titles, awards, etc) of a person shown in the infobox changes over time (Figure 3).

Concretely, for each biography, we pick its latest revision every X days since the first revision, where X is sampled uniformly from $[270, 450]$, to diversify the timestamps included in the data. We then extract the infobox and the lead paragraph per revision. As illustrated in Figure 3, two test sets are created: `test-same time` contains articles that are published at the same time as the training set, while `test-future` consists of samples that are created (or revised) after training and development sets. We further ensure that the subjects of biographies in both test sets are not in training or development sets. On average, each revision has 15.3 attributes in the infobox and 43.2 words in the first paragraph. Details of data collection are included in Appendix B.

5 Experiments and Results

For data-to-text generation on TEMPWIKIBIO, we linearize the infobox and use it as the input to BART. Common data-to-text metrics (Gehrmann et al., 2021) are used, including BLEU-4 (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), TER (Snover et al., 2006), and BERTScore (Zhang* et al., 2020). For text-to-text summarization on XSum (Narayan et al., 2018) that consists of news articles and their corresponding summaries from BBC News, we report ROUGE scores (Lin, 2004)

	Prompt	B-4 (↑)	MTR (↑)	TER (↓)	BS (↑)
Test Future	-	30.48	49.88	69.66	55.34
	ENC:T	30.76*	50.10*	69.33	55.38
	ENC:L	**31.22***	*50.32**	**68.66***	**55.79***
	DEC:T	*31.05**	**50.46***	69.98	55.35
	DEC:L	30.69*	49.94	*69.03**	*55.52**
Test Same Time	-	30.81	50.15	70.26	55.51
	ENC:T	31.27*	50.55*	*69.78**	*55.82**
	ENC:L	**31.50***	*50.56**	**69.33***	**56.00***
	DEC:T	*31.43**	**50.73***	70.51	55.62
	DEC:L	30.91	50.11	69.86	55.59

Table 1: Results on TEMPWIKIBIO. Textual (T) and linear (L) prompts are used on encoder (ENC) or decoder (DEC). B: BLEU; MTR: METEOR; BS: BERTScore. The best result per metric is in **boldface** and the second best is in *italics*. Improvement over the no-prompt baseline is shaded. *: significantly better than the baseline with approximate randomization test ($p < 0.001$).

	Content Transfer			XSum		
Prompt	R-L	MTR	BS	R-1	R-2	QEval
-	27.52	27.55	29.86	45.23	22.11	47.73
ENC:T	*28.15**	*28.40**	*30.59**	**45.63***	*22.38**	47.76
ENC:L	27.82*	27.99*	30.33*	45.32	22.22	47.76
DEC:T	**28.41***	**28.84***	**30.90***	*45.59**	**22.45***	47.70
DEC:L	27.62	27.68	30.13*	44.94	21.91	47.51

Table 2: Results on the content transfer and XSum datasets. R: ROUGE; QEval: QuestEval.

and QuestEval (Scialom et al., 2021), a QA-based faithfulness evaluation metric that checks if questions created from the summary can be addressed by reading the document with a QA model, and vice versa. The content transfer dataset (Prabhumoye et al., 2019) considers sentences containing citations of news sources in Wikipedia articles as the target for generation. Many target sentences incorporate important dates from the cited articles, thus making it suitable to test our time-aware prompt design. To generate each target sentence, the context passage, which contains three sentences preceding the target sentence, and the cited news article are provided as input. ROUGE-L, METEOR, and BERTScore are computed for evaluation on the content transfer dataset.

Automatic Evaluation. Overall, models with time-aware prompts obtain better performance than the no-prompt baseline. Models with time-aware prompts win 49 of all 56 comparisons against the baseline on different metrics and datasets, indicating that adding time-aware prompts encourages the models to generate more informative outputs.

Linear prompts tend to work better on the encoder when the input is structured data, achieving the best overall performance on TEMPWIKIBIO (Table 1). Linear prompts also show less performance degradation than textual prompts when testing future samples. However, the better extrapolation performance by the model with linear prompts might be due to its lower sensitivity to the provided dates, as later revealed in our analysis.

When the input and output are both in natural language, textual prompts are more suitable, as evidenced by the better performance than linear prompts on all metrics on both the content transfer dataset and XSum (Table 2). We think that on text-to-text generation tasks, textual prompts benefit from modal compatibility and have an advantage of connecting the salient content with the timestamps.

Human Evaluation. We hire three fluent English speakers to evaluate 80 sets of paragraphs generated for TEMPWIKIBIO samples at a future time and 80 sets of sentences generated for content transfer samples by two models with time-aware prompts. The judges compare the output by each model against the output by the no-prompt baseline on two aspects: **informativeness** — whether the model output covers salient information in the input; and **factuality** — whether the model output is factually correct. For each set, the judges only know which output is generated by the baseline, and outputs by other models are randomly sorted. Besides the three-way label (win/tie/lose), we ask the judges to determine if the difference in each aspect between each pair of comparisons is date-related.

As shown in Figure 4, only a small portion of improvements by linear prompts is date-related, especially on the content transfer dataset where none of the outputs by the model with linear prompts is rated as having better factuality due to the inclusions or changes of dates, suggesting that linear prompts might help process other content. By contrast, the model with textual prompts focuses more on the temporal information and brings more factual dates into the outputs on the content transfer dataset.

Analysis via Date Perturbation. We probe into date sensitivity, to understand the mechanisms behind the two types of prompts. Specifically, the original timestamps of 2000 samples randomly selected from the test set of each dataset are per-

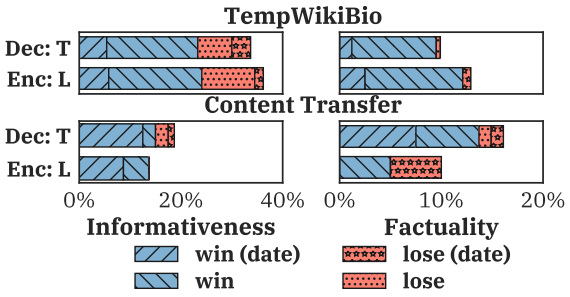

Figure 4: Percentages of samples that win or lose over the no-prompt baseline, on Test-Future of TEMPWIKIBIO and the content transfer dataset. While both time-aware prompts improve informativeness and factuality on TEMPWIKIBIO, textual prompts are more often rated as having date-related improvements in informativeness and factuality on the content transfer dataset. Krippendorff's α: 0.85 (informativeness); 0.64 (factuality).

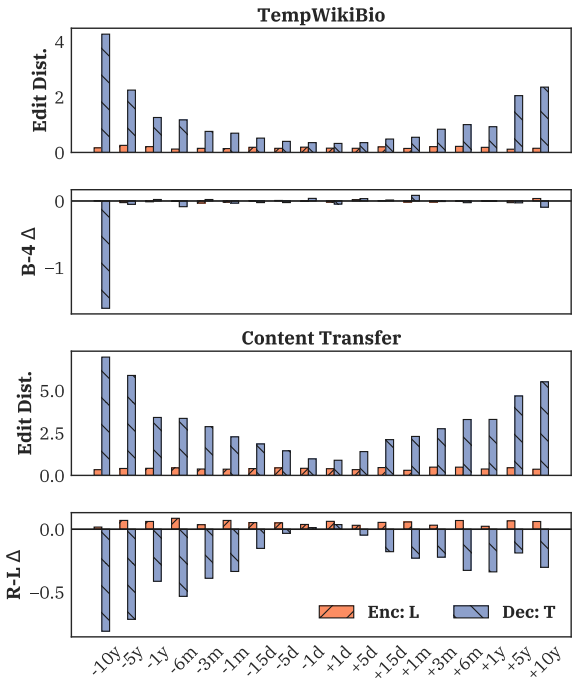

Figure 5: Edit distances and differences of BLEU-4/ROUGE-L between outputs with perturbed dates and original dates on TEMPWIKIBIO and content transfer. Linear prompts are not sensitive to the given dates. Results on XSum are in Appendix C.

turbed and provided to the models. As indicated by the greater edit distances between the outputs produced with the perturbed dates and original dates (Figure 5), models with textual prompts are more sensitive to the given dates than models with linear prompts. Human inspection of the outputs by the model with linear prompts and perturbed dates also finds that their changes from the original outputs are not related to the temporal information, echoing that the improvements in informativeness and factuality are less date-related according to human judges. For the model with textual prompts, we observe a need for learning complicated world knowledge to generate correct dates more frequently when provided perturbed dates, as shown in Figure 1 and Figure 9 in Appendix G.

In addition, the greater differences of ROUGE-L scores suggest a more significant dependency on the temporal information by the content transfer dataset, where the publication dates of documents are often required to generate the outputs (Figure 8 in Appendix G), calling for the inclusion of metadata during data collection.

6 Conclusion

We study two types of time-aware prompts for injecting document timestamps into generation models. Experiments on TEMPWIKIBIO, our newly collected data-to-text generation dataset, and two text-to-text generation tasks show that linear prompts mostly enhance the processing of content other than dates for more informative and factual outputs. Textual prompts build the association between the given temporal information and the generated temporal information, producing outputs with more factual dates.

Acknowledgements

This work is supported in part by National Science Foundation through grant IIS-2046016, and Oracle Cloud credits and related resources provided by the Oracle for Research program. We thank the anonymous reviewers for their valuable suggestions.

Ethical Consideration

Our work assumes the timestamps of documents can be accurately obtained and the models are always provided with the accurate creation dates. However, this might not be the case for some documents, especially the ones that are first published in a paper format and later digitized into electronic versions. Informing generation models of inaccurate timestamps could lead to incorrect content generation and other unpredictable behaviors, where fabricated facts might be picked up by end users, potentially causing harm to the public.

Limitation

Though we show that textual time-aware prompts help models generate more factually consistent outputs, we find that models with temporal prompts could generate incorrect temporal information due to the lack of world knowledge (Figure 9 of Appendix G). In this work, we do not further study methods that can incorporate extra world knowledge to address this issue.

During model evaluation, we investigate the effects of time-aware prompts on the generated temporal information via human evaluation, which includes 160 outputs by each model (320 in total). We believe automatic metrics that verify the correctness of temporal information in the outputs can better validate the improvements by our models. However, such automatic metrics do not exist. A potential design of temporal information evaluation metrics is to combine event and temporal expression extraction systems. We made several attempts at this design, but the performance of the event and temporal expression extraction systems we tested needs further improvement.

To obtain the timestamp of each sample, we rely on the automatic web archive (Appendix B). However, this approach for timestamp retrieval only applies to datasets that are based on web sources (e.g., news articles and blog posts). In addition, less popular web sources are less likely to be archived by automatic web archive service, which makes retrieving their timestamps more complicated and prevents the adoption of our methods.

References

Oshin Agarwal and Ani Nenkova. 2021. Temporal effects on pre-trained models for language processing tasks. *CoRR*, abs/2111.12790.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2021. Time-aware language models as temporal knowledge bases.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun KIM, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*.

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

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web*, pages 625–635.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34.

R'emi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

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

Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–71, Brussels, Belgium. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, null null, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Shrimai Prabhumoye, Chris Quirk, and Michel Galley. 2019. Towards content transfer through grounded text generation. 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 2622–2632, Minneapolis, Minnesota. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine

Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural DiversiTy on the Social Web*, DETECT '11, page 35–40, New York, NY, USA. Association for Computing Machinery.

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: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Prompt Design	R-1	R-2	R-L
1 (*Date: ...*)	45.70	**22.57**	37.47
2 (*Today is ...*)	**45.71**	22.55	**37.49**
3 (*The following ...*)	45.61	22.54	37.32

Table 3: ROUGE scores on the dev set of XSum by models with different textual prompts on the encoder. Textual prompt *"Today is [converted timestamp]."* achieves the highest average ROUGE score and is used in our main experiments.

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

A Details of Prompts

Other Textual Prompts. The three designs of textual prompts we try in our pilot study are:

1. *Date: [converted timestamp].*
2. *Today is [converted timestamp].*
3. *The following text is written on [converted timestamp].*

As shown in Table 3, the second design *"Today is [converted timestamp]."* yields the highest average ROUGE score on the development set of XSum. Note that performances by the three prompt designs do not vary greatly.

B Details of Datasets

B.1 TEMPWIKIBIO

Data Collection. We use the English Wikipedia dump[2] processed on February 1, 2022 and collect revisions before 2021 that have complete infoboxes. To identify biographies from all articles, we use the article titles and IDs in WIKIBIO (Lebret et al., 2016), which are originally from WikiProject for biographies.[3] We then extract attributes of the infobox and the first paragraph of the article from each remaining revision with mwparserfromhell.[4] Revisions that do not contain complete infoboxes are discarded. We further discard the first five revisions of each article to avoid including revisions with less comprehensive information about the person. To limit the number of revisions for each

[2] https://dumps.wikimedia.org/

[3] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Biography

[4] https://github.com/earwig/mwparserfromhell

Year	Train	Dev	Test-Same Time	Test-Future
2004	56	3	1	-
2005	2,167	118	10	-
2006	24,509	1,234	132	-
2007	76,637	3,794	421	-
2008	133,467	6,770	802	-
2009	194,504	9,882	1,209	-
2010	248,768	12,578	1,685	-
2011	309,642	15,801	2,266	-
2012	371,008	18,893	2,927	-
2013	365,260	18,619	3,039	-
2014	397,155	20,572	3,531	-
2015	441,800	22,597	4,286	-
2016	535,693	27,771	5,462	-
2017	460,233	23,733	4,602	-
2018	445,955	22,872	4,169	-
2019	-	-	-	11,698
2020	-	-	-	11,998
2021	-	-	-	11,132

Table 4: The numbers of TEMPWIKIBIO revisions made in different years, grouped by different splits.

article, we pick the latest revision every X days, where X is sampled uniformly from $[270, 450]$ after picking a revision to diversify the timestamps of selected revisions.

For the test split, the number of articles is down-sampled to 10% of the original number of articles created in the test time period to achieve a reasonable running time of decoding.

Time Statistics. The numbers of revisions made in different years are shown in Table 4.

Copyright Policy. We comply with the Wikipedia copyright policy[5] to collect the TEMPWIKIBIO. TEMPWIKIBIO will be released under the CC BY-SA 3.0 license.[6] The usage of TEMPWIKIBIO is limited by the copyright policy of Wikipedia.

B.2 Content Transfer

The content transfer dataset (Prabhumoye et al., 2019) extracts sentences with citations to news outlets in Wikipedia as the sentences to be generated. The cited source documents then become the external source documents where the generation should be grounded. The three sentences preceding each sentence to be generated in the original Wikipedia article are taken as the context passage. As the source documents come from many different news sources, instead of constructing an extraction template for each new source, we query the Wayback Machine [7] for the date when each source document was first archived to obtain the timestamp.

Year	Train	Dev	Test
unknown	1,719	14	250
1996	186	1	19
1997	100	1	8
1998	34	0	3
1999	399	3	29
2000	1,012	13	92
2001	831	11	99
2002	3,561	57	294
2003	6,672	150	553
2004	4,119	63	443
2005	4,112	66	441
2006	7,323	89	581
2007	12,398	151	937
2008	17,920	176	1,562
2009	26,090	298	2,499
2010	30,505	311	2,612
2011	40,035	400	3,256
2012	71,152	708	6,219
2013	94,216	799	7,274
2014	84,661	830	7,193
2015	55,248	588	5,229
2016	49,927	572	4,455
2017	44,243	509	3,489
2018	22,200	215	2,316
2019	1,337	22	147
Total	580,000	5,000	50,000

Table 5: Numbers of source documents for content transfer published in different years, grouped by different splits.

Time Statistics. In Table 5, we report the numbers of source documents in the content transfer dataset published in different years.

Copyright Policy. The content transfer dataset is publicly available[8] with the usage limited by the MIT License.[9]

B.3 XSum

We conduct experiments on text summarization with XSum (Narayan et al., 2018), which contains articles from BBC News. During the construction of the dataset, the first sentence of each article is taken as the summary of the remaining content. The timestamp of each news article is extracted from its corresponding HTML file.

Time Statistics. In Table 6, we report the numbers of articles in XSum published in different

[5] https://en.wikipedia.org/wiki/Wikipedia:Copyrights

[6] https://creativecommons.org/licenses/by-sa/3.0/

[7] https://web.archive.org

[8] https://www.microsoft.com/en-us/research/project/content-transfer

[9] https://opensource.org/licenses/MIT

Year	Train	Dev	Test
2009	0	0	1
2010	1,142	60	62
2011	2,820	154	153
2012	5,450	304	319
2013	7,939	409	420
2014	15,409	810	855
2015	49,041	2,792	2,736
2016	70,922	3,928	3,983
2017	51,322	2,875	2,805
Total	204,045	11,332	11,334

Table 6: Numbers of XSum articles published in different years, grouped by different splits.

	Prompt	Parent	# Date
Test Future	-	56.30	1.33
	ENC:T	56.40	1.33
	ENC:L	**56.57***	1.33
	DEC:T	*56.52**	1.35
	DEC:L	56.36	1.32
Test Same Time	-	57.57	1.25
	ENC:T	57.74*	1.25
	ENC:L	*57.82**	1.25
	DEC:T	**57.88***	1.26
	DEC:L	57.57	1.24

Table 7: Additional results on TEMPWIKIBIO. Textual (T) and linear (L) prompts are used on the encoder (ENC) or decoder (DEC). The best result is in **boldface** and the second best is in *italics*. Improvement over the no-prompt baseline is shaded. *: significantly better than the baseline with approximate randomization test ($p < 0.001$).

years.

Copyright Policy. XSum dataset is publicly available[10] with the usage limited by the MIT License.

C Additional Results

TEMPWIKIBIO. We additionally evaluate the model outputs on TEMPWIKIBIO with PARENT (Dhingra et al., 2019) and report the average number of dates in each model output. As shown in Table 7, the trend of PARENT scores is similar to other metrics, where the model with linear prompts on the encoder achieves the best result on samples drawn at a future time, while the model with textual prompts on the decoder achieves the best result on samples drawn at the same time period of the training and development sets.

Content Transfer and XSum. We report BLEU-4 on content transfer and ROUGE-L on XSum in Table 8. Textual prompts yield the best performance on the two datasets.

[10] https://github.com/EdinburghNLP/XSum

	Content Transfer		XSum	
Prompt	BLEU-4	# Date	ROUGE-L	# Date
-	11.05	0.530	37.04	0.275
ENC:T	*11.52**	0.617	**37.34***	0.289
ENC:L	11.27*	0.548	37.15	0.272
DEC:T	**11.66***	0.610	**37.34***	0.293
DEC:L	11.10	0.536	36.84	0.287

Table 8: Additional results on the content transfer and XSum datasets.

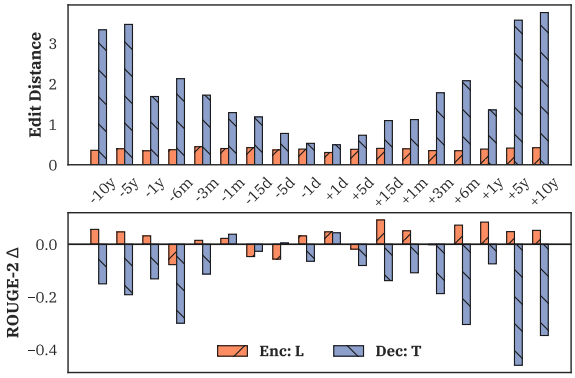

Figure 6: Edit distances and differences of ROUGE-2 between outputs with perturbed dates and original dates on XSum. Similar to the results on TEMPWIKIBIO and content transfer, linear prompts are not sensitive to the given dates.

Sensitivity Analyses. We show the results of sensitivity analyses with date perturbation on XSum in Figure 6. Linear prompts again do not show sensitivity to the given dates. Compared to feeding the model with dates that are a year later or earlier, greater drops of ROUGE-2 are observed when feeding the models with dates that are 6 months later or earlier. This suggests that XSum emphasizes resolving date relations of months and days.

D Experiments with T5

We also conduct experiments with the T5 pre-trained model (Raffel et al., 2019). As `T5-large` has 370 million more parameters than `bart.large` that has 400 million parameters, we use `T5-base`, which has 220 million parameters. We do not experiment on XSum, where the performance by T5 is shown to be much lower than BART (Gehrmann et al., 2021).

Results on TEMPWIKIBIO and content transfer are shown in Tables 9 and 10. While linear prompts still work better on TEMPWIKIBIO samples drawn

	Prompt	B-4 (↑)	MTR (↑)	TER (↓)	BS (↑)
Test Future	-	25.24	50.29	70.89	42.94
	ENC:T	25.19	50.27	71.03	42.86
	ENC:L	**25.25**	**50.32**	**70.79**	**42.96**
Test Same Time	-	24.33	50.00	72.30	42.54
	ENC:T	**24.38**	**50.06**	**72.19**	**42.58**
	ENC:L	24.35	50.01	72.21	42.57*

Table 9: Results on TEMPWIKIBIO with T5 as the base model. Textual (T) and linear (L) prompts are used on encoder (ENC) or decoder (DEC). B: BLEU; MTR: METEOR; BS: BERTScore. The best result per metric is in **boldface**. Improvement over the no-prompt baseline is shaded.

Prompt	R-L	B-4	MTR	BS
-	22.71	7.65	22.31	24.04
ENC:T	**23.17***	**8.00***	**23.11***	**24.73***
ENC:L	22.68	7.61	22.30	23.96

Table 10: Results on the content transfer dataset with T5 as the base model. R: ROUGE. *: significantly better than the baseline with approximate randomization test ($p < 0.001$).

at a future time and textual prompts work better on text-to-text content transfer, the improvements by linear prompts are less substantial. We conjecture that T5 is pre-trained with natural language prefixes for multiple tasks and prefers textual prompts.

E Details of Human Evaluation

Figures 10 and 11 include the instructions provided to annotators for our human evaluation. All annotators are college students based in the U.S. The purpose of the annotation study and the usage of collected data are explained to the annotators before the annotation begins. We compensate each annotator with $15 per hour.

F Details of Implementation

For experiments with BART (Lewis et al., 2020), we use `bart.large`.[11] For experiments with T5 (Raffel et al., 2019), we use `T5-base`.[12] Fairseq (Ott et al., 2019)[13] is used for model training and decoding with BART. HuggingFace Transformer (Wolf et al., 2020) is used for decoding with T5. Experiments are conducted with NVIDIA A6000 GPU with 48GB memory.

Training Settings. For training on all datasets with BART, we first follow the hyperparameter setting provided by the original BART training script for XSum[14] except that we set the total number of update steps to 30,000 for TEMPWIKIBIO and 35,000 for the content transfer dataset. In addition, we adjust the accumulated batch size for training on TEMPWIKIBIO to have 65,536 tokens in each batch. We then tune the learning rates on TEMPWIKIBIO and the content transfer dataset by searching through 1×10^{-5}, 3×10^{-5}, and 5×10^{-5} with the model without prompts. Based on the BLEU-4 scores on the development sets, we choose 5×10^{-5} for TEMPWIKIBIO and 3×10^{-5} for the content transfer dataset. Each model is trained for one run with one random seed due to the high computational cost of fine-tuning large models. For experiments with T5, we follow the default parameters suggested by HuggingFace.

Decoding Settings. We use beam search with beam sizes of 4, 4 and 6 for decoding on TEMPWIKIBIO, content transfer, and XSum. The maximum decoding lengths are set to 100, 100, and 60 for TEMPWIKIBIO, content transfer, and XSum.

Running Time. When using 4 GPUs, training on TEMPWIKIBIO, content transfer, and XSum takes 11, 7, and 2 hours. Meanwhile, decoding on TEMPWIKIBIO, content transfer, and XSum respectively takes 2 hours, 1 hour, and 15 minutes with 1 GPU.

Evaluation. We use sacreBLEU (Post, 2018)[15] for calculating the BLEU and TER scores. To obtain the METEOR (Lavie and Agarwal, 2007) score, we use NLTK (Bird et al., 2009). The official BERTScore (Zhang* et al., 2020)[16], QuestEval (Scialom et al., 2021)[17], and PARENT (Dhingra et al., 2019)[18] libraries are used. For ROUGE scores (Lin, 2004), we use the Python implementation by Google.[19]

[11] https://github.com/pytorch/fairseq/tree/main/examples/bart

[12] https://huggingface.co/t5-base

[13] https://github.com/pytorch/fairseq/tree/2380a6e4

[14] https://github.com/pytorch/fairseq/blob/main/examples/bart/README.summarization.md

[15] https://github.com/mjpost/sacrebleu

[16] https://github.com/Tiiiger/bert_score

[17] https://github.com/ThomasScialom/QuestEval

[18] https://github.com/google-research/language/tree/master/language/table_text_eval

[19] https://github.com/google-research/google-research/tree/master/rouge

G Example Outputs

In Figures 7, 8, and 9, we show example outputs by different systems on TEMPWIKIBIO, content transfer, and XSum.

Date: 2020-12-07 **Infobox Attributes:** name[Jan Hellström] birth_date[February 21, 1960] birthplace[Söderköping , Sweden] position[Striker] national_years[1986–1989]
Reference: Jan Hellström (born February 21, 1960 in Söderköping, Östergötland) is a former Swedish footballer, who played as a striker.
BART: Jan Hellström (born 21 February 1960 in Söderköping) is a former Swedish footballer.
Textual Prompt: Jan Hellström (born 21 February 1960 in Söderköping) is a former Swedish footballer.
Linear Prompt: Jan Hellström (born 21 February 1960 in Söderköping) is a Swedish former footballer who played as a striker.

Date: 2021-05-30 **Infobox Attributes:** name[George Daniel Brown] birth_date[1836] death_date[1902] birth_place[London , England] masters[DNP] open[5th: 1860 , 1863]
Reference: George Daniel Brown (1836–1902) was an English professional golfer. He also specialized as a ball and club maker. Brown finished fifth in the 1860 Open Championship and repeated the feat in the 1863 Open Championship. In total, Brown had three top-10 finishes in The Open Championship.
BART: George Daniel Brown (1836–1902) was an English professional golfer who played in the late 19th century. Brown had two top-10 finishes in the Open Championship. His best performance came in the 1860 Open Championship, when he finished in fifth place.
Textual Prompt: George Daniel Brown (1836–1902) was an English professional golfer. Brown placed fifth in the 1860 and 1863 Open Championships.
Linear Prompt: George Daniel Brown (1836–1902) was an English professional golfer. Brown placed fifth in the 1860 and 1863 Open Championships.

Date: 2021-02-08 **Infobox Attributes:** name[Robert A. King] coach_years1[1890] coach_team1[Wabash] overall_record[0–3]
Reference: Robert A. King was an American football coach. He was the fifth head football coach at Wabash College in Crawfordsville, Indiana, serving for one season, in 1890, and compiling a record of 0–3.
BART: Robert A. King was the third head college football coach for the Wabash College Little Giants located in Crawfordsville, Indiana and he held that position for the 1890 season. His career coaching record at WABash was 0 wins, 3 losses, and 0 ties. As of the conclusion of the 2007 season, this ranks him 18th at Wabat
Textual Prompt: Robert A. King was the third head college football coach for the Wabash College Little Giants located in Crawfordsville, Indiana and he held that position for the 1890 season. His career coaching record at WABash was 0 wins, 3 losses, and 0 ties. As of the conclusion of the 2007 season, this ranks him 18th at Wabat
Linear Prompt: Robert A. King was an American football coach. He served as the head football coach at Wabash College for one season in 1890, compiling a record of 0–3.

Figure 7: Example system outputs on TEMPWIKIBIO. Textual prompts are on the decoder and linear prompts are on the encoder.

Context: The second season of ”Faking It”, an American single-camera romantic comedy, premiered on September 23, 2014, and concluded on November 2, 2015, on the MTV network.
Date: 2014-06-17 **Source Document:** A day before its season finale, the comedy from showrunner Carter Covington has earned a sophomore run of 10 episodes. MTV is going to be Faking It for another year. The youth-skewing cable network has renewed the comedy starring Katie Stevens and Rita Volk, The Hollywood Reporter has confirmed. MTV will unspool 10 new episodes – up two from season one – in 2015. From showrunner Carter Covington, the half-hour comedy series was the first new scripted entry picked up to series under new network topper Susanne Daniels. "Faking It has proved to be the perfect companion show to Awkward, retaining nearly 90 percent of its lead in each week,” Daniels said in a release announcing the news Monday. “We’re excited about Carter Covington’s delicious plans for season two.” The comedy, which centers on two best friends who are mistakenly outed as lesbians and catapult to instant popularity, opened in April to 1.17 million total viewers. Through its first seven episodes, the comedy has averaged 948,000 total viewers. MTV says the show is the highest-rated new series launch this year with a 1.5 rating among viewers 12-34 and 1.4 million viewers each week when factoring in three days of delayed viewing. The season finale airs Tuesday. For MTV, Faking It comes as Daniels is looking to double the network’s roster of original scripted series. In addition to veterans Awkward and Teen Wolf, MTV will also launch comedy Happyland and dramas Finding Carter and Eye Candy. On the pilot side, MTV is readying its adaptation of Screamand has buzzy book adaptation Shannarain development.
Reference: On June, 2014, the series was renewed for a second season of 10 episodes, which was later extended to 20 episodes.
BART: The series was renewed for a second season of 10 episodes on September 23, 2014.
Textual Prompt: On June, 2014, MTV renewed ”Faking It” for a second season of 10 episodes.
Linear Prompt: The series was renewed for a second season of 10 episodes on September 23, 2014.

Context: Furthermore, the magazine said of the phrase, "So when your square friend uses it, take a little bit of pleasure in knowing they’re referencing a stoner comedy – or a drag reality show referencing a stoner comedy – even if they have no idea." In 2014, VH1 began airing a television show called ”Bye Felicia”, and pop singer Jordin Sparks released a mixtape titled ”#ByeFelicia”. According to Google Trends, the phrase reached its highest usage in mid-2015.
Date: 2017-12-15 **Source Document:** Outgoing White House official Omarosa Manigault Newman says Roberts’ ’Bye, Felicia’ dig was "petty" and a "black woman civil war." A link has been sent to your friend’s email address. A link has been posted to your Facebook feed. To find out more about Facebook commenting please read the Conversation Guidelines and FAQs Following an appearance on ’Good Morning America’ about her resignation from the White House, anchor Robin Roberts muttered the phrase about Omarosa Manigault Newmans. USA TODAY Omarosa Manigault Newman clapped back at Robin Roberts after the Good Morning America host took a dig at her on Thursday’s show, telling Inside Edition that her comments were "petty" and akin to "a black woman civil war." The White House public liaison and assistant to the president, who is leaving her post on Jan. 20, had told GMA’s Michael Strahan, "When I can tell my story — and it is a profound story — I know the world will want to hear." Later, a skeptical Roberts commented, "She said she has a story to tell? I’m sure she’ll be selling that story." Then she invoked Friday’s classic two-word dismissal for persons unlikely to be missed by anyone: "Bye, Felicia." Do you know how awful you have to be to annoy Robin Roberts? pic.twitter.com/h0gxAhRFRD For anyone unfamiliar with the expression, we’ll let its inventor, Ice Cube, explain: "It’s the phrase to get ANYBODY out of your face that’s saying something stupid."
Reference: On December 14, 2017, Robin Roberts, a host of ABC TV’s ”Good Morning America”, used the phrase to conclude a segment about Omarosa Manigault Newman’s departure from the Presidency of Trump administration staff.
BART:In January 2017, Omarosa Manigault Newman responded to Robin Roberts’ use of the phrase on ”Good Morning America”, calling it a "black woman civil war".
Textual Prompt: In December 2017, ”Good Morning America” host Robin Roberts used the phrase during an interview with Omarosa Manigault Newman about her resignation from the White House.
Linear Prompt: In January 2017, ”Good Morning America” host Robin Roberts used the phrase to refer to Omarosa Manigault Newman, who had just resigned from her position in the White House.

Figure 8: Example system outputs on the content transfer dataset. Textual prompts are on the decoder and linear prompts are on the encoder. The publication dates are frequently required in the outputs.

Date: 2016-07-04 **XSum Document:** The cloning of the first animal from an adult cell was a remarkable scientific achievement. It promised new treatments for debilitating diseases. But it also raised fears of cloned human beings, designer babies and a dystopian future. Twenty years on, neither the hopes nor the fears have been realised. So what is Dolly's legacy? I first saw Dolly in 1997 at the Roslin Institute just outside Edinburgh. She stood apart from the other sheep in the pens at this agricultural research centre. She stood prouder, her fleece seemed like a lion's mane and there was an aura about her. Dolly's creation had echoes of Mary Shelley's classic novel Frankenstein, in which inanimate tissue was brought to life by electricity. Dolly was created from DNA taken from a cell taken from an sheep. The technique involved putting the DNA into an empty eggshell and then zapping it with electricity. This created an embryo. Researchers at Roslin then implanted the embryo into the womb of a sheep which grew into Dolly - an exact genetic copy of the sheep from which the skin cell was taken. It took 277 attempts to clone Dolly and there were many miscarriages on the way. There were also genuine concerns that it would not be long before cloned humans would be walking the Earth - people would try to clone themselves to achieve a kind of immortality or they might try to resurrect a beloved dead relative. The airwaves were filled with conversations about what it meant to be human, whether the clones would be exactly the same as the person from which they were cloned and what kind of world the scientists were tumbling us into. When I met the researchers at Roslin they were acutely aware of public suspicion. And they knew it was important to be clear, open and honest about their work. Dolly's creator, Prof Sir Ian Wilmut, could not be any more different from fictional scientists such as Dr Frankenstein or indeed Dr Moreau, who developed human-like hybrids in a story by HG Wells. By contrast, Prof Wilmut is a quietly-spoken man whose motivation was to improve animal welfare and human health. This bespectacled scientist with his gentle candour was ideally complemented by the enthusiastic and eloquent Dr Simon Best, who was brought in to commercialise the technology to improve human health ...
Reference: The birth of Dolly the sheep seemed one of those moments in scientific research that would change the world forever.
BART:The birth of Dolly the sheep in 1997 sent shockwaves through the scientific community.
Textual Prompt: It is 20 years since the birth of Dolly the sheep.
Linear Prompt: It is 20 years since the birth of Dolly the sheep.

Date: 2016-04-05 **XSum Document:** Sunderland's Rachael Laws, 25, has been called up to replace the 32-year-old Liverpool keeper, who has a knee injury. England face Belgium in Rotherham on Friday, before travelling to Zenica to play Bosnia on Tuesday. The Lionesses have won both their qualifying games so far. Goalkeepers: Karen Bardsley (Manchester City), Rachael Laws (Sunderland), Carly Telford (Notts County) Defenders: Laura Bassett (Notts County), Lucy Bronze (Manchester City), Gilly Flaherty (Chelsea), Alex Greenwood (Liverpool), Steph Houghton (Manchester City), Alex Scott (Arsenal), Casey Stoney (Arsenal), Demi Stokes (Manchester City), Amy Turner (Notts County) Midfielders: Katie Chapman (Chelsea), Jordan Nobbs (Arsenal), Jo Potter (Birmingham City), Jill Scott (Manchester City), Fara Williams (Arsenal) Forwards: Eniola Aluko (Chelsea), Karen Carney (Chelsea), Gemma Davison (Chelsea), Toni Duggan (Manchester City), Fran Kirby (Chelsea), Ellen White (Notts County).
Textual Prompt: Manchester City goalkeeper Karen Bardsley has been ruled out of England's Euro 2017 qualifiers against Belgium and Bosnia-Herzegovina.
Perturbed Date: 2012-04-05 **Textual Prompt**: Manchester City goalkeeper Karen Bardsley has been ruled out of England's Euro 2012 qualifiers against Belgium and Bosnia-Herzegovina.
Perturbed Date: 2011-04-05 **Textual Prompt**: Manchester City goalkeeper Karen Bardsley has been ruled out of England's Euro 2012 qualifiers against Belgium and Bosnia-Herzegovina.
Perturbed Date: 2021-04-05 **Textual Prompt**: Manchester City goalkeeper Karen Bardsley has been ruled out of England's Euro 2021 qualifiers against Belgium and Bosnia-Herzegovina.
Perturbed Date: 2020-04-05 **Textual Prompt**: Manchester City goalkeeper Karen Bardsley has been ruled out of England's Euro 2020 qualifiers against Belgium and Bosnia-Herzegovina.

Figure 9: Example system outputs on XSum for text summarization. Textual prompts are on the decoder and linear prompts are on the encoder. In the first example, Dolly the sheep was actually born on July 5, 1996. In the second example, the Women's Euro is held every four years. Therefore, it could only be Euro 2013, 2017, or 2021.

Annotation Instruction

The annotation task consists of **80** groups of paragraphs produced by **two** systems that briefly describe the career of a person. In addition to the paragraphs, each group includes a infobox listing out important information about the person. You will also find a reference paragraph and a baseline paragraph for each group.

Please read each system-produced paragraph and compare it with the baseline paragraph on two aspects: **informativeness** and **factuality**. For each aspect, if the system-produced paragraph is better, please label "win"; if the system-produced paragraph is worse, please label "lose"; if the two paragraphs are similar, please label "tie".

When you label "win" or "lose", if the better or worse aspect is due to date mentions, please label "win (date)" and "lose (date)" correspondingly.

The explanation of the two aspects is shown below along with an example.

Example

Vincent Trapp

Personal information	
Born	26 January 1861 Melbourne, Australia
Died	21 October 1929 (aged 68) Melbourne, Australia
Domestic team information	
Years	**Team**
1881-1884	Victoria
	Source: Cricinfo, 23 July 2015

Baseline: Vincent Trapp (26 January 1861) was an Australian cricketer. He played two first-class cricket matches for Victoria between 1881 and 1884.
System1: Vincent Trapp (26 January 1861 – 21 October 1929) was an Australian cricketer. He played two first-class cricket matches for Victoria between 1881 and 1884.
System2: Vincent Trapp (26 January 1861 – 21 October 1929) was an Australian cricketer. He played for Victoria between 1881 and 1884.
System3: Vincent Trapp (26 January 1861) was an Australian cricketer. He played two first-class cricket matches for Victoria between 1881 and 1884, according to Cricinfo.

Informativeness: Whether the paragraph synthesizes salient information about the person.
In this example, both system 1 and 2 are better than the baseline as they mention the death date of the person, which is an important information, while system 3 that additionally talks about the source of the information ties with the baseline. Moreover, system 1 and system 2 should be labeled with "win (date)" as the information is related to a date.

Factuality: Whether the content of the paragraph is factually correct.
In this example, both system 1 and 3 tie with the baseline. System 2 is better than the baseline as it avoids mentioning "two first-class cricket matches" which is an incorrect information. System 2 should only be labeled with "win".

Figure 10: Guideline for human evaluation on WIKIREVISION.

Annotation Instruction

The annotation task consists of **80** groups of sentences produced by **two** systems that continue the given context passage using the information in the given source documents. In addition to the sentences, each group includes the context passage and the source document. You will also find a reference sentence and a baseline sentence for each group.

Please read each system-produced sentence and compare it with the baseline sentence on two aspects: **informativeness** and **factuality**. For each aspect, if the system-produced sentence is better, please label "win"; if the system-produced sentence is worse, please label "lose"; if the two sentences are similar, please label "tie".

When you label "win" or "lose", if the better or worse aspect is due to date mentions, please label "win (date)" and "lose (date)" correspondingly.

The explanation of the two aspects is shown below along with an example.

Example

Context Passage: The Burleigh Waters Library opened in 1991. For decades a local urban myth maintained that sharks were seen as far south in the canal waterways as Burleigh Waters. Alleged sightings and stories were locally spread, but balanced with scepticism.

Source Document: Publication date: 20 February 2003. The Queensland government has warned people not to swim in coastal canal systems after the second fatal shark attack in as many months on the Gold Coast yesterday. An 84-year-old man from Burleigh Waters died after he was attacked by a 2.5 metre bull whaler while swimming in Burleigh Lake just before 6.30am (AEST) ...

Baseline: An 84-year-old man from Burleigh Waters was attacked by a 2.5 metre bull whaler while swimming in Burleigh Lake.

System1: An 84-year-old man from Burleigh Waters died after he was attacked by a 2.5 metre bull whaler while swimming in Burleigh Lake.

System2: In February 2003, an 84-year-old man from Burleigh Waters died after he was attacked by a 2.5 metre bull whaler while swimming in Burleigh Lake.

System3: In 2013, an 84-year-old man from Burleigh Waters died after he was attacked by a 2.5 metre bull whaler while swimming in Burleigh Lake.

Informativeness: Whether the sentence synthesizes salient information of the source document.

In this example, all systems are better than the baseline as they mention the man died after the attack, which is an important information. Moreover, system 2 should be labeled with "win (date)" as it also mentions the date of the event.

Factuality: Whether the content of the sentence is factually correct.

In this example, both system 1 and 2 tie with the baseline. System 3 is worse than the baseline as it mentions an incorrect date. System 3 should be labeled with "lose (date)".

Figure 11: Guideline for human evaluation on the content transfer dataset.