Mapping FATE Concerns for Text Summarization

Su Lin Blodgett
Microsoft Research

Jackie C.K. Cheung
Mila / McGill University

Alexandra Olteanu
Microsoft Research

Adam Trischler
Microsoft Research

Abstract

The applications of textual summarization range from storyline generation and sentence compression to email commitment reminders, among others. Despite this breadth, we still have a poor understanding of the range of FATE (fairness, accountability, transparency and ethics) concerns that could arise from text summarization tasks and their applications. This is especially problematic if the summary is meant to replace the source materials, as users would be oblivious to misrepresentations caused by the summarizer. Could the generated summaries erase contributions? Could they misgender individuals? Are there variations in the quality of the summaries across different groups that either wrote the original text or are referenced in it? To address this gap, the goal of this project is two-fold: 1) construct a typology of FATE concerns for text summarization and 2) construct an inventory of diagnosis tests for these concerns.

1 Introduction and Background

Approaches for summarizing text typically either select and copy relevant fragments of the source text (extractive approaches) or select and paraphrase such fragments (abstractive approaches) [Kryściński et al., 2019a]. Models are often evaluated via automatic metrics that measure lexical overlap between generated and gold standard summaries. A more recent line of work aims to ensure that generated summaries are consistent with the source text, as abstractive systems risk generating so called "hallucinations," i.e., text that distorts content in the original document or is factually incorrect [Cao et al., 2020, Dong et al., 2020, Falke et al., 2019, Kryściński et al., 2019b, Kumar and Cheung, 2019].

No work has yet comprehensively examined the FATE concerns that arise from summarization systems, despite their increasing use and the clear risks of generating incorrect or harmful summaries. For example, we speculate that generated summaries might misgender the people they describe; they might produce representational harms by emphasizing different qualities of the people they describe; they could give rise to libelous representations by failing to appropriately qualify or hedge claims, or otherwise mislead users by giving rise to inferences that are ambiguous or unsupported by the source text; represent contested topics unfairly; provide offensive or politically charged framings; or be susceptible to adversarial perturbations of the source text. Despite this wide array of potential concerns, little work has emerged in this space. The few examples include: Dash et al. [2019] examining if multi-document summaries fairly represent document authors belonging to distinct social groups. In restricted domains like opinion summarization, there has also been work on whether a summary reflects the distribution of opinions in the source documents [Carenini and Cheung, 2008].

Through this project we aim to contribute the first typology mapping out the landscape of potential FATE concerns arising from summarization systems. This typology will address both the shared challenges that arise across different summarization tasks and the challenges that arise for specific applications, providing a research agenda for future work in responsible summarization.

Preprint. Under review.

In addition to identifying FATE concerns, a complementary aim is to identify potential sources and mechanisms of these concerns by examining the interplay between source datasets, external data, models, tasks, and usage contexts. This will help us develop appropriate measurements grounded in application contexts for a range of FATE concerns. For example, evaluations of the stereotyping produced by generated summaries need to account for the information present in source texts that constrains the summaries, requiring new metrics. Artificial perturbation tests and other adversarial approaches [Prabhakaran et al., 2019, Sato et al., 2018, Li et al., 2020, Rajagopal et al., 2020] could also help surface such scenarios for a range of harms.

2 Methodology and Deliverables

The goal of our exploratory project is two-fold:

1) *Construct a typology of FATE issues for summarization.* Doing so requires understanding what the standard evaluation approaches for text summarization *do* account for, and how current inventories of computational harms and FATE concerns apply to text summarization applications.
   - The typology will be grounded in a **text summarization literature survey** to examine what issues have been considered, including the quality criteria along which summaries are evaluated.
   - These quality criteria will then be contrasted with existing **inventories of computational harms for language technologies** [Bender, 2019, Blodgett et al., Olteanu et al., 2020].
2) *Construct diagnosis tests for a range of FATE issues.* Concomitantly, to demonstrate how the typology can inform new evaluation frameworks, for existing abstractive or extractive summarization applications we will define measurements of e.g., representational harms, as well as adapt and develop datasets for measurement and evaluation purposes.

2.1 A typology of FATE concerns for text summarization

We aim to assemble the first comprehensive typology that maps the landscape of FATE issues for summarization. Such a typology should necessarily include both shared challenges across different summarization tasks and challenges that arise for specific tasks and applications. We will ground the project primarily in applications where short summaries are generated for assistive purposes, such as document headlines [Tan et al., 2017] and summary TODOs from email threads [Mukherjee et al., 2020]. This avoids the complexity of evaluating long-text generations [Celikyilmaz et al., 2020] and attendant challenges to diagnostic development.

To develop this typology we will start with a *survey of the text summarization literature* to examine the work that has emerged across NLP, ML, and AI venues, along with quality criteria used to assess generated summaries. Drawing on similar analyses of natural language generation evaluation [Celikyilmaz et al., 2020] and "bias" in NLP [Blodgett et al., 2020], we will systematically survey what criteria are accounted for by current evaluation methods, what assumptions (about users, deployment contexts, etc.) underpin these criteria and methods, and what FATE concerns the existing criteria may address. The quality criteria and deployment settings identified across the literature will be contrasted against inventories of computational harms for language technologies [Bender, 2019, Blodgett et al., Olteanu et al., 2020] to identify arising concerns that a) are covered by existing inventories but not by summarization quality criteria, and b) are specific to summarization and not covered by existing inventories.

2.2 Development of measurements and diagnosis tests for FATE concerns

Guided by the results of our typology of FATE concerns above, we plan to develop measurements and diagnosis tests that can provide concrete measures of progress. Because whether a generated summary gives rise to a FATE concern often depends on the content of the underlying source text, the summarization setting presents a challenge for developing valid and reliable measurements that also capture what may be normatively desirable. For example, a summary may only be able to describe people of different genders in an appropriately similar fashion (e.g., with their titles and careers) if that information is available in the source text.

Inventory of existing datasets and development of new diagnosis datasets. We will start by cataloguing existing summarization datasets and examine how they might be adapted for measuring FATE concerns; these datasets include CNN/Daily Mail [Hermann et al., 2015], XSum [Narayan et al.,

2018], MultiNews [Fabbri et al., 2019], Newsroom [Grusky et al., 2018], NYTimes [Sandhaus, 2008], and Curation [Curation, 2020]. Since these were not developed with specific FATE concerns in mind, however, many might not be fit for our purpose. Some FATE concerns, like measuring quality-of-service harms [Blodgett et al., 2020] in the absence of demographic data, will thus require developing new diagnosis datasets that account for how demographic cues are embedded in language. To assemble such datasets, we will employ a mix of data sampling, perturbation, and crowdsourcing techniques [Olteanu et al., 2020, Fabbri et al., 2019, Prabhakaran et al., 2019].

Exploration of underlying mechanisms. To generate effective diagnosis tests, the project will examine the interplay between source datasets, external data, models, tasks, and usage contexts to characterize the mechanisms underlying measured concerns. For example, we might examine the relative contributions of source datasets and various models (e.g., BART [Lewis et al., 2019] and Pegasus [Zhang et al., 2020]) to undesirable representations of different people. Concretely, one direction we propose is *adversarial perturbations*—minimal nudges to model inputs that induce significant undesired changes in their outputs. Generating these perturbations automatically for language input is challenging, since language is discrete, and often comes at the expense of interpretability. However, recent advances have improved the situation [Sato et al., 2018] and we have shown how to leverage automated word swaps and substitutions to robustify NLP models for common sense [Emami et al., 2019]. We also plan to use crowdsourcing here to verify the quality of the perturbations.

2.3 Project Impact

Automatic summarization and the technologies that rely on it are important to Microsoft and other industry stakeholders. For example, Office has a feature that distills key points from users' Word documents. For such a feature to be acceptable to users, let alone useful and successful, it must adhere to standards for Responsible AI. This feature and summarization tools like it are still in their infancy. Deployed summarization systems have often been extractive only—the safer but less natural option. A major blocker to the deployment of abstractive summarization is "hallucination": the generation of text that distorts content in the original document or is factually incorrect. An audit of common and harmful distortion types, as we propose here, will be an important first step in remedying them. We believe this project has the potential to set a long-term, mixed-methods, cross-disciplinary research agenda. Our survey, analysis, and metric innovations will not only lay the foundation for research that builds effective, responsible summarization systems, but also have immediate product impact.

3 Resources and Budget

The project will be driven by two graduate students, who will be jointly supervised by the MSR and Mila co-PIs. The Mila co-PI, Jackie Cheung, will be actively involved in the project and directly supervise the students at Mila. From MSR Montréal, Su Lin Blodgett, Alexandra Olteanu and Adam Trischler will be actively involved in mentoring, including, e.g., weekly group meetings and 1:1s with the students.

We plan for one student to focus on the typology construction part of the project (Section 2.1), and one to focus on how to operationalize an inventory of FATE issues in the context of text summarization applications and construct corresponding diagnosis tests (Section 2.2). We envision submitting at least two publications on each of these complementary directions. Thus, we plan to support two full-time graduate students at MILA on this project. The standard stipend for PhD students at Mila is $25,000 for one year. Their tuition will be covered by other funding sources. In addition, we budget $2,500 per student per year for conference travel and printing costs, assuming that international conferences will resume next year. Given the dataset construction component of the project, we will also run a series of crowdsourcing experiments which we estimate to require an $8,000 budget (Section 2.2). Additional crowdsourcing resources will be covered through the MSR FATE MTL crowdsourcing budget, if needed. Thus, our request for the project is $63,000.

Research Team

Su Lin Blodgett is a postdoctoral researcher at Microsoft Research Montréal with the Fairness, Accountability, Transparency and Ethics (FATE) group.

Jackie C.K. Cheung is an assistant professor and Canada CIFAR AI Chair at the Mila Quebec AI Institute. He has worked extensively on automatic summarization with a focus on evaluation methodology and factual correctness in abstractive summarization systems.

Alexandra Olteanu is a principal researcher at Microsoft Research Montréal, part of the Fairness, Accountability, Transparency and Ethics (FATE) group.

Adam Trischler is a principal research manager at Microsoft Research Montréal, and the lead of the Deep Learning & Language group.

References

Emily M. Bender. A typology of ethical risks in language technology with an eye towards where transparent documentation can help, 2019. Presented at The Future of Artificial Intelligence: Language, Ethics, Technology Workshop. `https://bit.ly/2P9t9M6`.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language and Justice. In preparation.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *ACL*, 2020.

M. Cao, Y. Dong, J. Wu, and J. C. K. Cheung. Factual Error Correction for Abstractive Summarization Models. In *EMNLP*, 2020.

G. Carenini and J. C. K. Cheung. Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proc. of INLGC*, 2008.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of Text Generation: A Survey. *arXiv preprint arXiv:2006.14799*, 2020.

Curation. Curation corpus base, 2020.

A. Dash, A. Shandilya, A. Biswas, K. Ghosh, S. Ghosh, and A. Chakraborty. Summarizing User-Generated Textual Content: Motivation and Methods for Fairness in Algorithmic Summaries. *Proc. ACM Hum.-Comput. Interact.*, (CSCW), 2019.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. Multi-fact correction in abstractive text summarization. In *EMNLP*, 2020.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proc. of ACL*, 2019.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *ACL*, 2019.

Tobias Falke, Leonardo F.R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *ACL*, 2019.

M. Grusky, M. Naaman, and Y. Artzi. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *NAACL*, 2018.

K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching Machines to Read and Comprehend. In *NIPS*, 2015.

W. Kryściński, N. S. Keskar, B. McCann, C. Xiong, and R. Socher. Neural Text Summarization: A Critical Evaluation. In *EMNLP*, 2019a.

W. Kryściński, B. McCann, C. Xiong, and R. Socher. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*, 2019b.

K. Kumar and J. C. K. Cheung. Understanding the Behaviour of Neural Abstractive Summarizers using Contrastive Examples. In *NAACL*, 2019.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. Contextualized Perturbation for Textual Adversarial Attack. *arXiv preprint arXiv:2009.07502*, 2020.

S. Mukherjee, S. Mukherjee, M. Hasegawa, A. H. Awadallah, and R. White. Smart to-do: Automatic generation of to-do items from emails. *arXiv preprint arXiv:2005.06282*, 2020.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *EMNLP*, 2018.

A. Olteanu, F. Diaz, and G. Kazai. When Are Search Completion Suggestions Problematic? *Proc. ACM HCI*, (CSCW), 2020.

V. Prabhakaran, B. Hutchinson, and M. Mitchell. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In *EMNLP*, 2019.

Dheeraj Rajagopal, Niket Tandon, Bhavana Dalvi, Peter Clark, and Eduard Hovy. What-if I ask you to explain: Explaining the effects of perturbations in procedural text. *arXiv preprint arXiv:2005.01526*, 2020.

Evan Sandhaus. The New York Times Annotated Corpus LDC2008T19, 2008. Philadelphia: Linguistic Data Consortium.

M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto. Interpretable adversarial perturbation in input embedding space for text. In *IJCAI*, 2018.

J. Tan, X. Wan, and J. Xiao. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *IJCAI*, 2017.

J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *ICML*, 2020.