Responsible AI Considerations in Text Summarization Research

A Review of Current Practices

Yu Lu Liu[1, 2], Meng Cao[1, 2], Su Lin Blodgett[3], Jackie Chi Kit Cheung[1, 2, 4], Alexandra Olteanu[3], Adam Trischler[3]

[1]Mila – Quebec Artificial Intelligence Institute [2]McGill University
[3]Microsoft Research, Montréal, Canada [4]Canada CIFAR AI Chair

Motivation & Overview

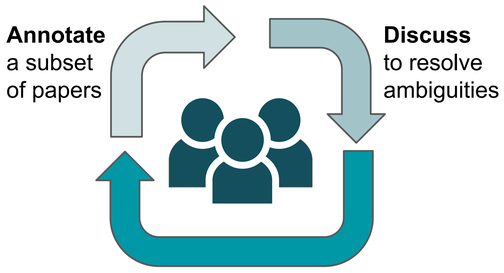
For the task of automatic text summarization, our understanding of how prevalent responsible AI (RAI) issues are, or when and why these issues are likely to arise, remains **limited**.

We investigate how, when, and which RAI issues are covered in the contemporary text summarization literature:

- We develop a set of **annotation guidelines**
- We conduct a **systematic review** of >300 summarization papers

Annotation Scheme & Process

Multiple Rounds
To build and refine the annotation scheme



Annotation Scheme

Paper Authors & Goals
- Author affiliation
- Type of contribution
- Intended domain
- Research goal

Data & Evaluation Practices
- Data domain (actual domain)
- Evaluated quality criteria

Limitations & Ethical Considerations
- Limitations of prior work
- Limitations of one's work
- Ethical considerations
- Mentioned stakeholders

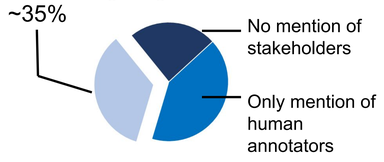
ACL Anthology 2020-2022 333 papers

Annotators
Each paper is annotated by one of us, or a graduate student in NLP.

Findings

How do practitioners describe the intended use contexts of their contributions?

- Many contributions are intended to be **general-purpose**:
  - ~55% of papers contributing systems
  - ~72% of … metrics
  - ~23% of … datasets
- Papers **seldom mention stakeholders** when imagining intended use contexts:



- Imagined benefits to anticipated users often only include:
  - **Reducing labor** (e.g., reduce workload by summarizing meetings)
  - **Improving customer experiences** (e.g., improve shopping experience by summarizing product reviews)

Intended use contexts are often not well-described

How is a "good" summary conceptualized?

- **Information saliency** (e.g., *"relevance," "informativeness," "redundancy"*): ~41% of all reviewed papers.
- **Linguistic properties** (e.g., *"coherence," "fluency"*): ~39%
- **Factuality** (e.g., *"factual consistency," "hallucination"*): ~28%

Criteria such as bias and usefulness are rarely evaluated.

What are common evaluation practices?

- **Mismatch between intended and actual domain**: ~52% of "general-purpose" systems *only* use news data in training/testing.
- **Heavy reliance on ROUGE-like metrics:**
  ~90% of systems use these metrics.
  ~22% of all papers *only* use them.

Current evaluation practices may not provide meaningful insights about systems' true performance

How do practitioners discuss limitations and ethical considerations of their work?

- **Most papers do not include such discussions:**

- **When authors conceptualize ethical concerns, they often turn to data-related issues.** However, **data bias** remains poorly defined or under-specified (e.g., data may contain *"biased views"* without further elaboration)
- Various quality criteria are discussed in limitations, but they are rarely also conceptualized as ethical concerns → only **factuality** is conceptualized as an ethical concern.
- Discussion of stakeholders is often limited to:
  - Compensation of human annotators
  - Data privacy
  - Intended positive impacts on anticipated users.

  → **Potential harm to stakeholders overlooked**

Authors engage with a narrow range of potential ethical concerns

Who are the practitioners?

Author Affiliation	#
Academic	299
Industry	121
(collab of above two)	(95)
Other	32

What kind of work do practitioners prioritize?

Type of Contribution	#
System (models, methods)	224
Dataset	91
Metric	36
Evaluation	73
Application & Other	34

Recommendations

We encourage practitioners to…

- Clearly articulate relevant stakeholders, intended domains, and potential impacts to those stakeholders.

- Consider using more stakeholder-centric quality criteria (e.g., bias, fairness, usefulness).
- Develop and adopt eval. practices tailored to specific use contexts.

- Reflect on the intended use context and on what is a "good" summary in that intended use context.
- Engage with prior literature on ethical concerns and harms in NLP.