Mapping FATE concerns for text summarization

Students: Jules Barbe (undergraduate), Meng Cao (graduate)

Researchers: Su Lin Blodgett, Jackie Cheung, Alexandra Olteanu, and Adam Trischler

FATE concerns in NLP

There Is a Racial Divide in Speech-Recognition Systems, Researchers Say

Computer says no: Irish vet fails oral English test needed to stay in Australia

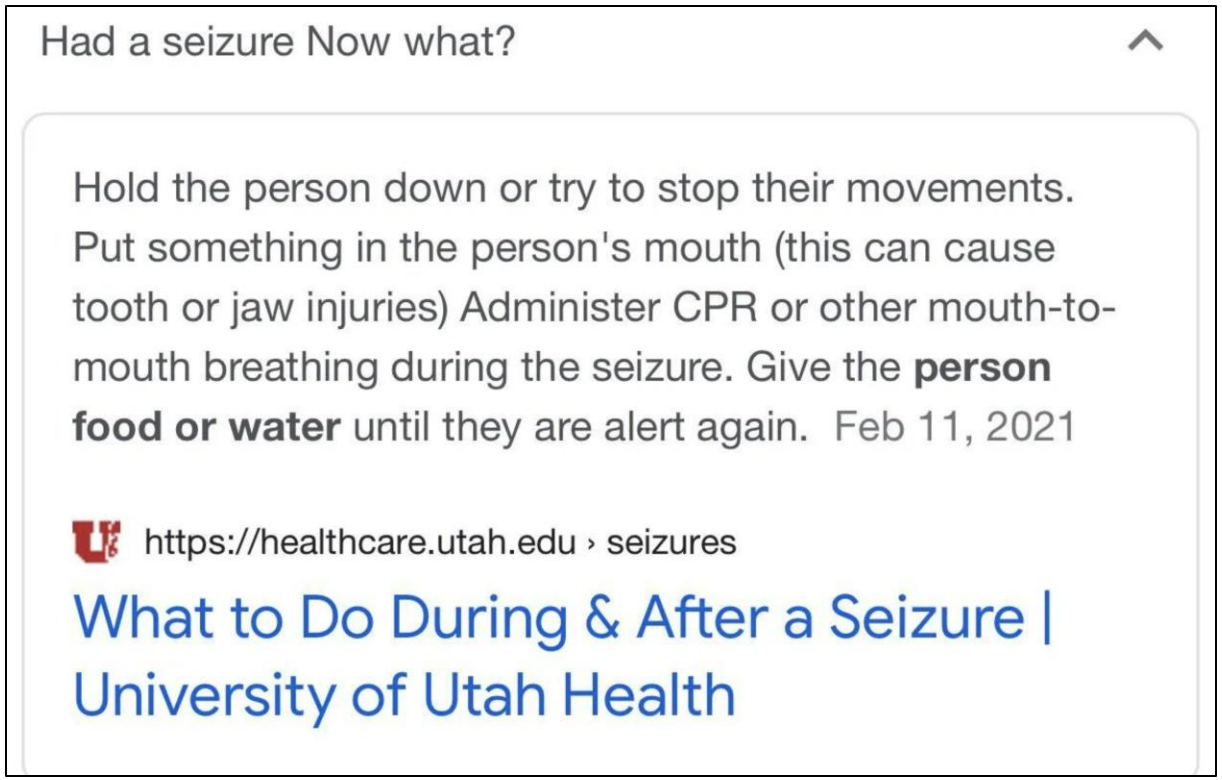
‘He works, She cooks’: Google Translate results reveal gender bias in tech

AI’s Islamophobia problem

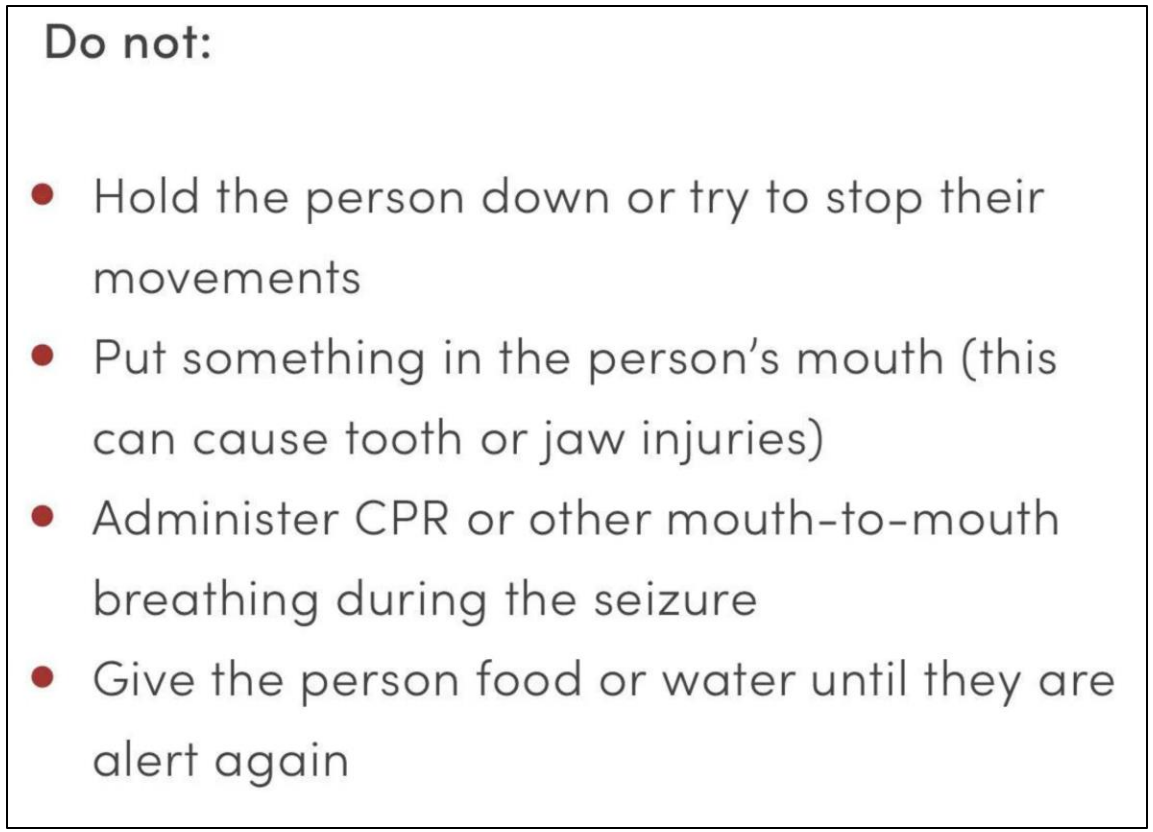
GPT-3 is a smart and poetic AI. It also says terrible things about Muslims.

FATE concerns in text summarization: **Misinformation**

Search summary

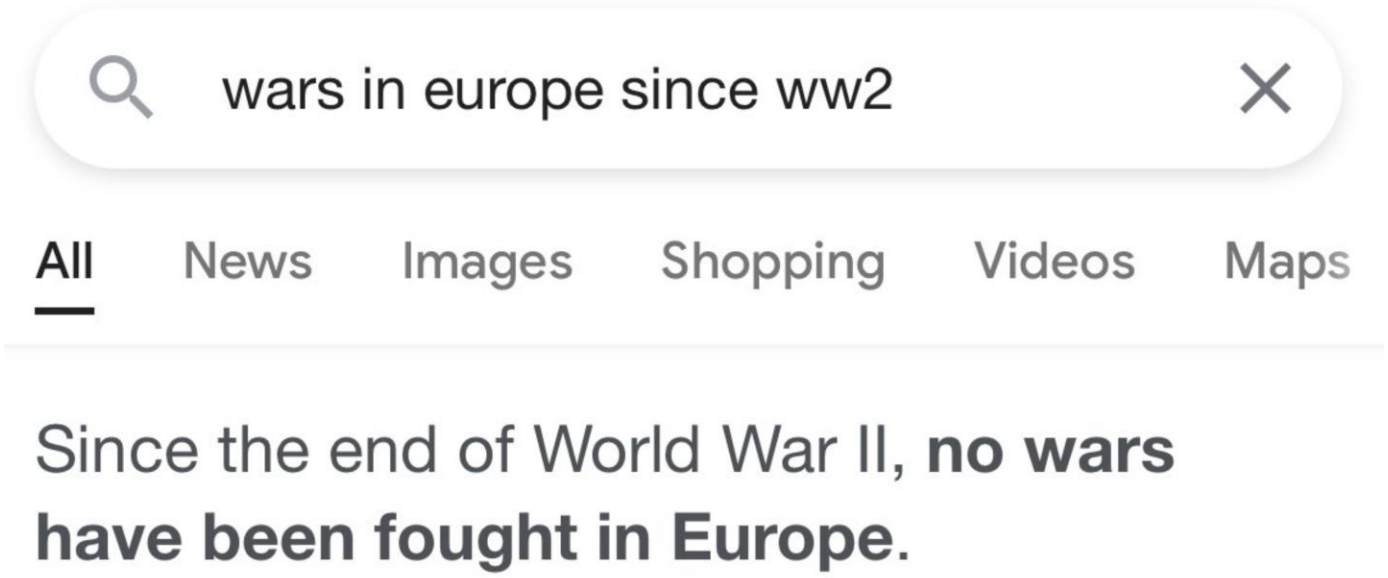


Source document



https://twitter.com/soft/status/1449406390976409600

FATE concerns in text summarization: **Erasure**



https://twitter.com/astrobassball/status/1498005840539885576

✗  HARMS MAPPING: HOW DO WE KNOW WHAT WE SHOULD BE MEASURING?

✗  HARMS MEASUREMENT: HOW SHOULD WE BE MEASURING HARMS?

✓  HARMS MITIGATION: HOW CAN WE PREEMPT & MINIMIZE HARMS?

For more see:

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. **Frontiers in Big Data. 2019.**

Language (Technology) is Power: A Critical Survey of "Bias" in NLP. S. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. **ACL 2020.**

Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Dataset. S. Blodgett, G. Lopez, A. Olteanu, R. Sim, and H. Wallach. **ACL 2021.**

Text summarization: Types & domains

Extractive summarization: selecting snippets of text to form summary

Abstractive summarization: synthesizing and rewriting source text to form summary

- Much more popular in the past 5 years with seq-to-seq-based models
- Potential to be more expressive, but also more harmful?

Domains:

- **News summarization** still predominant; many large corpora and benchmarks
- Less researched: scientific articles, opinions, conversations, books, dialogue, biographies, legal texts, arguments, medical documents

Project goals

Harms mapping

Construct a typology of FATE concerns for text summarization

Harms assessment

Construct an inventory of diagnosis datasets for these concerns

Challenges: Harms mapping

Stakeholders	stakeholder role stakeholder vulnerability stakeholder agency demographic cues
System affordances	restricted access unfair exposure unwarranted disclosure
System behavior & response	*what*: opportunity, resource allocation, dignity, representation, agency, autonomy, physical/emotional well-being *how*: immediately, frequently, by altering beliefs, by nudging towards an action *why:* harmful processes vs. outcome

Biega et al. (2018), Abbasi et al. (2019), Crawford (2017), Barocas et al. (2017)

Challenges: Harms measurement

What **observable properties** to use

Construct **validity**: Do we measure what we think we're measuring?

- Could the measurement be (inadvertently) capturing something else?
- Does the measurement have diagnostic utility?
- Could the measurement cause harm or shift incentives?

Reliability: Can the measurement be repeated?

Measurement and Fairness. A.Z. Jacobs and H. Wallach. [**FAccT '21**]

The current landscape

Work addressing FATE concerns is scarce, addressing only **hallucinations** and **fairness in representation** in multi-document summarization

Work rarely explicitly anticipates **stakeholders** or **deployment scenarios**

- ethical considerations sections tend to address resources used and how human evaluation was conducted
- increasing acknowledgment that systems may produce biased outputs, but explicit focus on these issues is less common

Evaluation criteria (content coverage, fluency) and accompanying **metrics/datasets** may not be well-equipped to capture FATE concerns

Preliminary: Taxonomy of harms

Summarization mechanisms

What summarization mechanisms might result in harmful outcomes?

Harmful outcomes

What types of possibly harmful outcomes do we observe as a result of using text summarization?

Others

Descriptors of issues (e.g., hallucinations), goals (e.g., changing tone)

Preliminary: Overview of mechanisms

Deletions

of modifiers
of identity markers
of clause markers
[…]

Out-of-vocabulary insertions

terms additions
terms replacement
[…]

In-vocabulary manipulations

text shuffling
terms preservation rates
[…]

Preliminary: Examples of possible harmful outcomes

Erasure harms

e.g., dialects, identity, topics

Stereotyping harms

e.g., by introducing term associations, or preserving stereotypical statements at higher rates

Misinformation harms

e.g., resulting statements are factually incorrect or misleading

Offensive speech

e.g., resulting text can be construed as offensive

[Other: privacy harms, promotes violence, encourage self-harm, framing bias, etc.]

Mapping mechanisms to harms

Mechanism	Possible types of harm	Example
Deletions		
Modifiers	Misinformation	**In:** The **alleged criminal** was seen... **Out:** The **criminal** was seen…
Clause markers	Misinformation	**In:** …**if** newly revised NYPD training materials are approved by a federal judge, new cadets could be taking courses… **Out:** …new NYPD training materials are approved by a federal judge…
Identity markers	Erasure (of identity) Framing bias	**In:** Police said the attackers – three **white** men and one **Asian** man – were racially motivated. The victims were **black**. Protesters have been gathered since 25-year-old Jamar Clark was shot during a struggle with police… **Out:** Four men have been charged with attempted murder after shots were fired at protesters…
Topics	Erasure (of perspectives), misinformation	**In:** Robert Kennedy Jr has apologized for describing the number of children injured by vaccines as 'a holocaust'….The film purports that there is a connection between thimerosal…and a rise in autism…**despite the majority of the scientific community dismissing any connection**. **Out:** Robert Kennedy Jr used the term last week during the screening of a film….[H]e publicly retracted his statement….

Mapping mechanisms to harms

Mechanism	Possible types of harm	Example
Out-of-vocabulary manipulations		
Terms replacement	Misinformation (defamation)	**In:** The 21 year old was charged with the **death** of… **Out:** The 21 year old was charged with the **murder** of…
Terms addition	Stereotyping	**In:** The kids were playing on the ground when a stone fell from the old building. **Out:** The **angry** kids were playing when the stone fell.
In-vocabulary manipulations		
Text shuffling	Misinformation (illogical)	**In:** Prince George could be days away from **becoming an older brother** as the duchess is due to give birth to her second child mid-to-late April. **Out:** Prince George is due to give birth in mid-to-late April **to an older brother**.
…		

How did we surface these examples?

Exploratory: Mapping mechanisms to harms

Literature review

to examine known issues and resulting harms

Qualitative explorations (examples/application driven)

to both reproduce known issues and uncover additional issues and resulting harms using available datasets and models

Quantitative explorations (e.g., deletion patterns)

to uncover possible harms due to systematic patterns

Literature review

Categorized 200+ papers according to:

- input domain
- summarization pipeline "stage" targeted (dataset, model, processing, evaluation)
- approach (abstractive, extractive or hybrid)
- input type (single or multi-document)
- language target (mono-, multi- or cross-lingual)
- output type (headline, highlights, concept map, full summary)

Some findings:

- Large focus into hallucinations when it comes to summarization issues.
- Possible issues through less known mechanisms such as erasure, entity-swapping, language variation, evaluation or dataset bias.
- Lack of transparency when it comes to application domain or possible use cases. Makes it harder to surface potential issues before model deployement.

Qualitative explorations

Generated 10 000+ summaries using recent summarizers and current benchmark datasets (Samsum for dialogue, CNN/DM and XSUM for news)

Especially targeted texts that contain words of interest (race, gender, religion)

Possible issues identified by manual analysis:

- Stereotyping
- Deletion
  - e.g., omission leading to erasure harm

Example:

In: A news article reporting on a violent attack during a protest, where the police describes the event as racially motivated against the black victims.

Out: The summary omits mentions of race.

(Dataset: XSUM, Model: PEGASUS, see slide 15)

Quantitative explorations: Word deletion

Drop rate: how often is a word present in a source text but *not* in the system-generated summary?

Preliminary observations:

- Disparity between negatively and positively connotated words
- (e.g., "friendly" with a 0.7 drop rate versus "violent" with a 0.81 drop rate using PEGASUS on XSUM)
- Some demographically sensitive words are frequently dropped (e.g., "Islamic" dropped in 28 out of 29 appearances)

Plan:

- Perturbation testing to check for fairness in dropping (e.g., replace "Islamic" with "Catholic" or "atheist")
- Assess how word deletions affect factuality of summary
- Control for saliency of words in original source document

Next directions

Time Period	Task
Now – end of March 2022	Finish qualitative analyses and drop rate experiments
Now – April 15, 2022	Finish paper and submit to ACL Rolling Review in mid-April
April 2022 – May 2022	Initial study focused on harms from erasure • Identify extent of problem in existing summarizers • Constructing samples where conflicts in opinions or information states must be resolved
June – Dec 2022	Build diagnostic evaluation and dataset on this issue • Systematic expansion of previous step into dataset and evaluation • **Goal**: release evaluation to research community