MMERB

A Multimodal Dataset for Measuring Marginalized Ethnicity Reporting Bias

Tom Södahl Bladsjö

January 1, 2024

Reporting Bias

"[...]the tendency of people to not state the obvious"
(Paik, Aroca-Ouellette, Roncone, & Kann, 2021)

The frequency with which people write about actions, outcomes, or properties is not a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.
(Chang, Ordonez, Mitchell, & Prabhakaran, 2019)

Gricean Maxim of Quantity:

- Make your contribution as informative as is required (for the current purposes of the exchange).
- Do not make your contribution more informative than is required.

Gricean Maxim of Relevance:

- Be relevant.

(Grice, 1975)

Reporting bias can seriously impact what a model trained on text learns about the world (Gordon & Durme, 2013; Paik et al., 2021).

A simple example

Bananas

Brown bananas

A simple example

Bananas

Brown bananas

Reasonable. However, humans are not bananas

Two examples from the Flickr8k dataset

A little girl in a pink dress going into a wooden cabin.

An asian girl in a pink dress is smiling whilst out in the countryside.

All humans are human...

...but some humans are more normative than others

...attributes that are not the norm in a certain setting will be mentioned more often than attributes that are normative (and therefore too obvious to mention)

Expression	Count	Expression	Count
black man	35	asian man	89
asian woman	28	asian woman	84
asian girl	28	black man	50
asian man	24	asian girl	32
asian girls	20	asian people	24
asian women	17	white man	23
asian boy	17	asian women	19
asian children	10	asian men	18
dark skinned man	9	white woman	15
white man	8	indian man	13

(a) Flickr8k ethnicity expression counts

(b) MS COCO ethnicity expression counts

Table: The number of times each expression appears in each dataset, ordered by frequency. Only the ten most common expressions are included.

Creating the dataset

The original datasets

Flickr8k

- 8092 images
- 5 captions for each image
- Images collected from Flickr
- Annotatated by crowd workers Amazon Mechanical Turk (Hodosh, Young, & Hockenmaier, 2013)

MS COCO (2017 train set)

- 118287 images (in the train set)
- Multiple annotations per image
- Also collected from Flickr
- Annotated by crowd workers at Amazon Mechanical Turk (Lin et al., 2014)

The MMERB dataset

- 1313 images
- Each image has two contrasting captions
- Created using simple pattern matching with regular expressions

Example

A little girl in a pink dress → A little white girl in a pink dress
An asian girl in a pink dress → A girl in a pink dress

Resulting dataset:

- Two sets of images:
  - 680 images of non-white people (test group)
  - 633 images of white people (norm group)
- Two captions for each image:

  One that mentions ethnicity...

  ...and one that does not
- Total of four test sets:
  - test_mention (680 instances)
  - test_no_mention (680 instances)
  - norm_mention (633 instances)
  - norm_no_mention (633 instances)

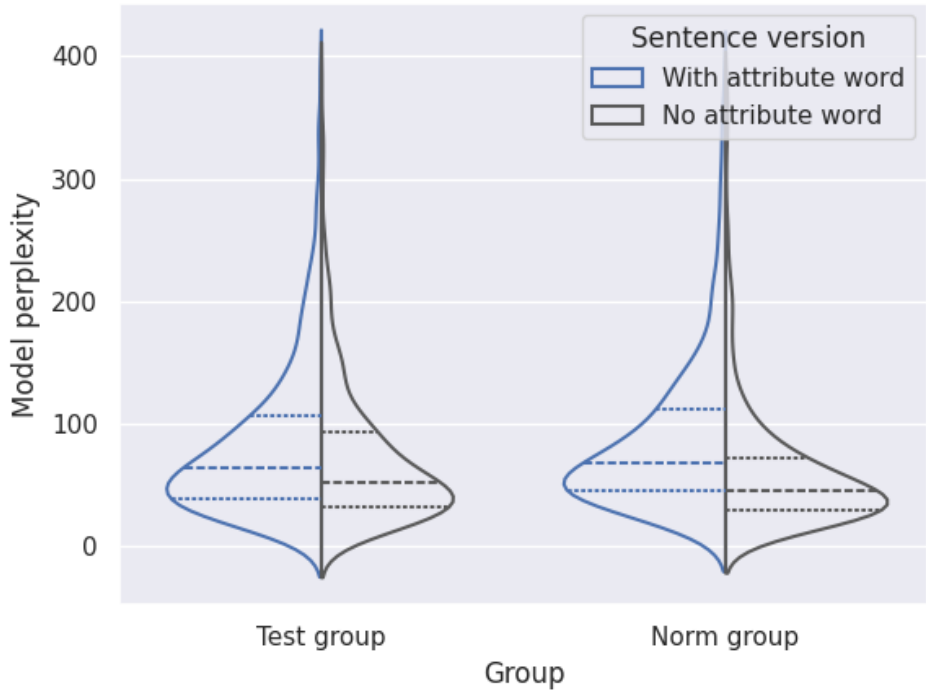
Intuition:

- The model is generally going to be less likely to mention ethnicity than to not mention it
- ...but *how* unlikely it is to mention ethnicity will vary depending on the ethnicity in question
- Hypothesis: The model will be less likely to describe a white person as white than to describe a non-white person with their ethnicity

How to use this dataset

- Follow the instructions for downloading and preparing the dataset at `github.com/TomBladsjo/LT-Resources-project`
- Test a multimodal model on each of the datasets (using a metric of your choice), keeping track of the order of the examples
- Compare model performance pairwise for the two captions on each image
- Compare these differences across groups

Result for BLIP for Conditional Generation[1] using perplexity as a test measure:



Difference in group means as measured by Welch's t-test: t-statistic = 11.4, p-value=1.13e-28

[1] https://huggingface.co/docs/transformers/model_doc/blip

Thank you!

ppt template credit:
`https://github.com/Urinx/LaTeX-PPT-Template`

Chang, K.-W., Ordonez, V., Mitchell, M., & Prabhakaran, V. (2019). *Tutorial: Bias and fairness in natural language processing.* Recorded presentation at EMNLP 2019, hosted at UCLA NLP `http://web.cs.ucla.edu/~kwchang/talks/emnlp19-fairnlp/`.

Gordon, J., & Durme, B. (2013, 10). Reporting bias and knowledge acquisition. In (p. 25-30). doi: 10.1145/2509558.2509563

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (p. 41 - 58). Leiden, The Netherlands: Brill. Retrieved from `https://brill.com/view/book/edcoll/9789004368811/BP000003.xml` doi: 10.1163/9789004368811_003

Hodosh, M., Young, P., & Hockenmaier, J. (2013, may). Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, *47*(1), 853–899.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*. Retrieved from https://api.semanticscholar.org/CorpusID:14113767

Paik, C., Aroca-Ouellette, S., Roncone, A., & Kann, K. (2021, November). The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 823–835). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.63 doi: 10.18653/v1/2021.emnlp-main.63