Inducing Positive Perspectives with Text Reframing

Caleb Ziems, Minzhi Li, Anthony Zhang, Diyi Yang

Adhitya Venkatraman, Joel Johnson
CS224C

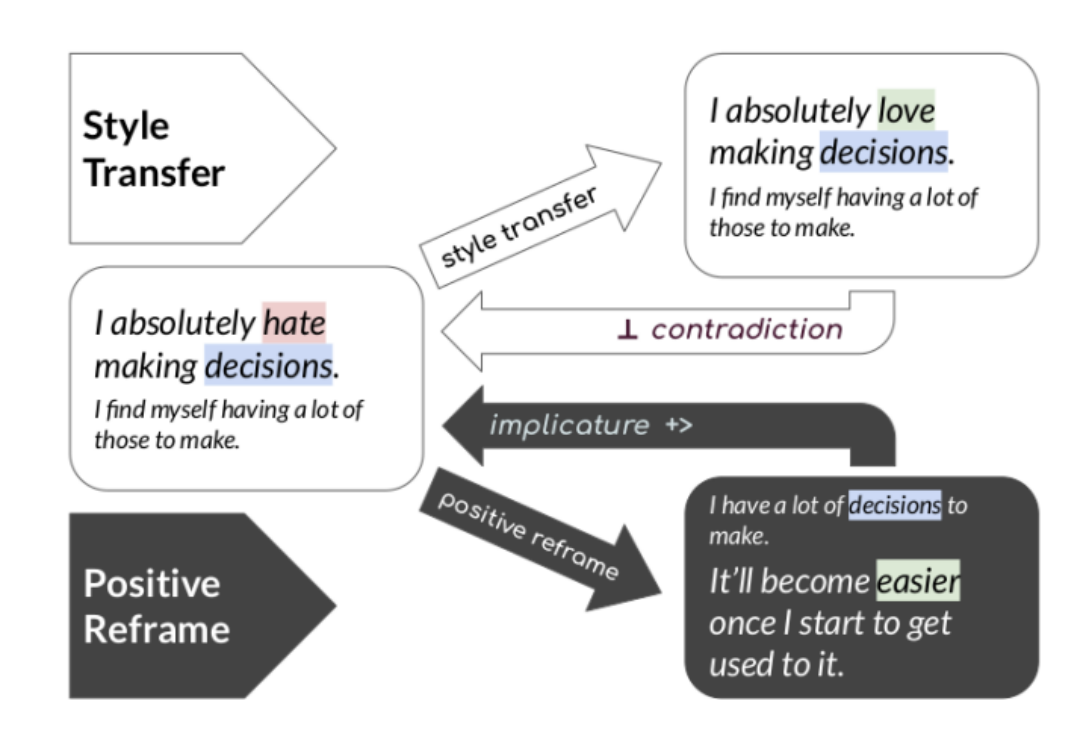
Roadmap

1. Live Demo
Reframing, ChatGPT vs. Humans, Evaluation

3. Additional Topics
Implications, Censorship, Connections to Other Papers

2. The Paper
Methods, Findings, Discussion

Positive Reframing vs. Style Transfer



Live Demo

"[It's possible] that although individuals from clinical groups can use reappraisal successfully when cued, they fail to appropriately identify moments at which ER would be helpful in everyday life" (McRae, Kateri, Gross)

Let's try reframing with ChatGPT!

Think about how comfortable you feel with the way that your sentence is reframed!

Does it still convey the same message? What aspects are gained or lost?

Demo

1. Suggest some phrases to reframe

2. We'll (qualitatively) evaluate the reframes on meaning, sentiment, and fluency

Reframing: Exercise & Discussion

- Now in groups, let's do this same exercise using human reframers.
- Are there certain topics where text reframing is allowed and others where it is not allowed?
- Which do you prefer, and when would you trust a model vs. a person/editor?

Methods and Experiment

- >1 million tweets filtered using the `#stressed` hashtag (2012-2021)
  - Notable processing steps: Offensive & overtly positive tweets excluded
- Crowdsourced annotators manually generated reframes (2 per tweet), scored the before/after
- Positive Reframing Strategies: Growth Mindset, Impermanence, Neutralizing, Optimism, Self-Affirmation, Thankfulness
- Three model settings: Unconstrained, Prediction, and Controlled increasingly involve the strategies
- Output is evaluated across Meaning, Positivity, and Fluency

Sample Reframes

Original Text	Strategy	Positive Reframe
(a) Could someone just give me like $1000? It would change my life, stressing about rent, bills and food money is just the worst.	growth	I need to learn how to manage my money better so that I don't have to stress about rent, bills and food money. *Time to focus on making a budget, so I don't have to stress about rent, bills, and food money.*
(b) I just went back to school today And I'm already stressed cause we have MIDTERMS NEXT WEEK !!! AND THIS WEEKEND – JAM PACKED :-(	impermanence	I just went back to school today and I'm already stressed because we have midterms next week and this weekend – JAM PACKED *Well, just think by the end of next week we'll be all done with the midterms!*
(c) talacre beach really need to sort there checking in routeen caravans are never ready 30mins later and still sat in the car waiting ..	neutralizing	The caravans at Talacre beach need to sort out their check-in process, so the kids don't get stuck in the car. *Talacre beach's check in process could do with some improvement to reduce the wait time for customers.*
(d) So glad that tomorrow is Friday. This has seriously been the longest week of my life! tired	optimism	I'm glad that tomorrow is Friday. This week has been long, but I'm looking forward to the weekend. *I'm glad the weekend is coming up, so I can rest.*
(e) Sometimes I get these impulses to just throw a tantrum. Like throw/break things, cry and scream.	self-affirmation	Sometimes I get these impulses to just throw a tantrum. Like throw/break things, cry and scream. But I'm strong, and I know I can handle it. *It's normal for to feel overwhelmed sometimes but I know I am strong to handle and go through it.*
(f) The fact that my mom's been calling me every night to calm me down from school says so much. I've needed it these past few days	thankfulness	I'm really lucky to have such a caring mum who is willing to call me every night when I'm having a hard time. *My mom has been calling me every night to calm me down from school. I've needed it these past few days. I'm thankful for her.*

...settling in on Twitter data using the "#stressed" parameter is bold. This is the first source of critique. **The period of data collection is so wide-spanning and behaviors so greatly evolved** in terms of the earnestness used in deploying hashtags**, I wonder if that influences the data that make it into the final dataset.**

- Allex Desronvil

However, since there is **only one suffix sentence annotated**, as shown in later sections of the paper, it is unsure that **how much meaning is preserved and if the remedy sentence is what the model users want.**

Another concern regarding the work is the hardness to evaluate...A faithful evaluation should include **how much content is preserved not only from the sense of similar words but also what events in the context are discussed**.

- Xinran Zhao

Generalizability

- Based on the methodology, how generalizable do you think the experiment is?
- What are the extensible vs. restraining factors? For example, consider the data source, processing methods, and theoretical literature.
- Ignoring methodological restrictions, what are the comparative advantages of analyzing long-form vs. short-form text?

Full Model Results

		Model	Automatic Evaluation R-1	R-2	R-L	BLEU	BScore	Δ TB	Avg. Len	Human Evaluation Meaning	Positivity	Fluency
Retrieval		Random	9.6	**3.6**	8.4	0.17	84.8	**0.36**	**20.0**	2.79	3.03	3.60
		SBERT	**15.2**	1.9	**12.8**	**1.47**	**87.6**	**0.36**	17.7	**3.45**	**3.97**	**4.16**
Few-shot		GPT-3	18.3	**3.4**	15.5	2.9	**88.2**	**0.44**	17.3	**3.73**	**4.17**	**4.27**
		GPT-Neo	**18.7**	**3.4**	**16.0**	**3.0**	**88.2**	0.40	**17.6**	3.69	4.16	4.21
Unconstrained	$p(t\|s)$	GPT	13.3	1.8	11.3	1.1	86.4	0.37	21.1	3.55	3.91	4.08
		GPT-2 No-pretrain	13.2	1.3	11.4	0.66	**89.6**	0.37	16.9	3.11	3.66	3.96
		GPT-2	20.9	4.6	17.7	4.2	88.5	0.35	20.0	3.58	4.01	**4.18**
		Seq2Seq-LSTM	15.7	1.4	12.4	0.73	85.6	**0.49**	25.8	3.33	4.15	4.10
		CopyNMT	20.8	5.0	18.0	4.0	85.7	0.32	16.1	3.57	3.69	3.91
		T5	27.4	9.8	23.8	8.7	88.7	0.38	**35.3**	4.09	3.79	4.06
		BART	**27.7**	**10.8**	**24.3**	**10.3**	89.3	0.23	24.4	**4.13**	**3.81**	4.15
Predict	$p(t, \psi_t\|s)$	T5	**27.5**	**10.5**	24.0	**11.0**	89.0	0.23	**25.1**	**4.10**	3.64	**4.11**
		BART	27.3	10.2	**24.1**	9.85	**89.4**	**0.32**	23.4	4.09	**3.95**	**4.11**
Control	$p(t\|s, \psi_t)$	T5	27.7	10.0	23.9	8.8	88.8	**0.36**	**35.0**	4.11	3.89	4.07
		BART	**28.8**	**10.9**	**25.1**	**10.1**	**89.6**	0.27	24.7	**4.23**	**4.07**	**4.27**
		Human	*100*	*100*	*100*	*100*	*100*	*0.35*	*17.4*	*3.80*	*3.82*	*4.18*

Table 2: **Positive reframing results** meausred by Meaning including ROUGE-1 (R-1), ROUGE-1 (R-2), ROUGE-L (R-L), BLEU, BERTScore (BScore), Positivity via Δ TextBlob (Δ TB) and Fluency. State-of-the-art models can generate meaning-preserving reframes in the unconstrained setting $p(t|s)$ and strategy-predictive setting $p(t, \psi_t|s)$ as well as when we condition the generation to use the reframing strategy from the ground truth $p(t|s, \psi_t)$. The best in-category performance is **bolded**; best overall performance is **highlighted**.

Key Findings

- T-5 and BART were the top performing models
  - GPT performed the worst; pre-training decisions can affect performance
  - Few-shot GPT-3 and GPT-Neo underperformed supervised models
  - BART generates a natural reframe *and* preserves meaning the best!
- Supervised Models Outperformed Simple Retrieval
- Error Analysis of 100 Randomly Sampled Model Generations
  - 26% contained *insubstantial changes*
  - 9% were *contradictions to the premise*
  - 6% were *self-contradictions*
  - 2% were *hallucinations*
- *Frame Strategy is Learnable*: *Growth, Impermanence, Neutralizing, Optimism, Self-Affirmation, Thankfulness*

I think the hallucinations is pretty funny but am most interested in how 1-3 may be combatted and if **fine tuning to eliminate these errors will make these models harder to transfer to other languages or across different cultures**.

- Clare Chua

The error analysis presents some issues to applying the current models to this subject, as the researchers explain. The **insubstantial changes seen in 26% of the generations could be particularly harmful as individuals may feel that their negative emotions are reinforced.** I think avoiding "contradictions to the premise" errors might be challenging for very negative framings—since **it might be hard to put a positive spin on something if someone has declared that nothing good is possible**, etc.—which is unfortunate as these are probably the statements that could benefit from positive framing the most.

- Sarah Bitter

Error Analysis

- Generally speaking, how do you feel about the error rate for the 100 example subset?

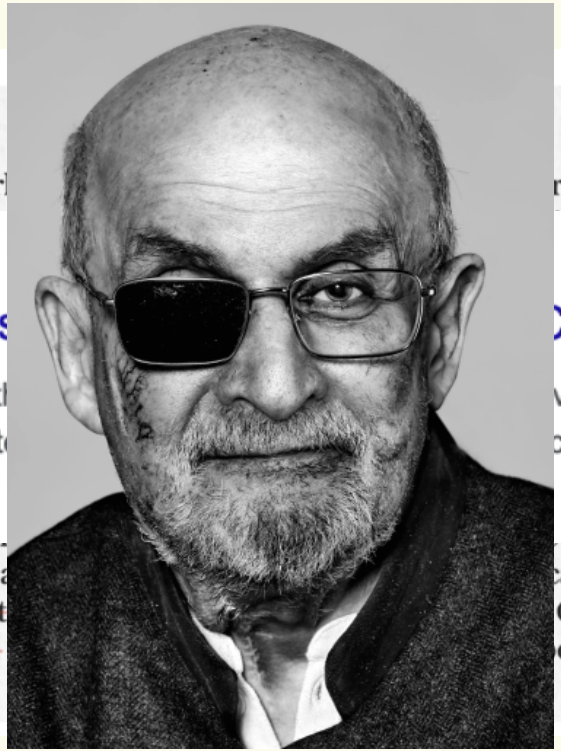
- Can you think of some examples of generations that may be difficult to classify in the following error classes? E.g. “insubstantial” criteria not generalizing across communities of practice?
    - Insubstantial Change
    - Contradictions to Premise
    - Self-Contradiction

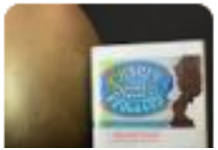
Positive Reframing Applications

- What are some applications of positive reframing technology?

- For example, consider:
  - Mental health and wellness
  - Hospital information delivery
  - Training and education
  - Social media moderation
  - News media
  - Digital assistants
  - Public relations
  - Market signals

Censorship

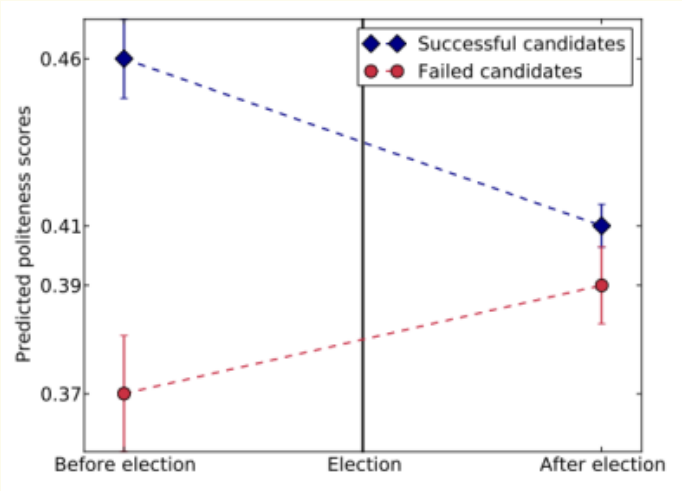




Censorship

- Recall that the BART model was able to preserve meaning in many instances, but edge cases seem really important here!
- Should others be allowed to reframe someone else's speech, even if it makes the text more inclusive, less offensive, etc.?
- What are some norms that could govern when and where we use this technology to adjust speech?

Politeness as Social Currency



- Last week: "A computational approach to politeness with application to social factors."
- How much do we buy their argument from a theoretical perspective?
- Can positive text reframing give people an undue advantage in achieving positions? Do we care if it does?

Thanks!