Civil Rephrases Of Toxic Texts With Self-Supervised Transformers

Léo Laugier[1], John Pavlopoulos[2,3], Jeffrey Sorensen[4], Lucas Dixon[4]

[1]Télécom Paris, Institut Polytechnique de Paris
[2]Athens University of Economics and Business
[3]Stockholm University
[4]Google

EACL 2021

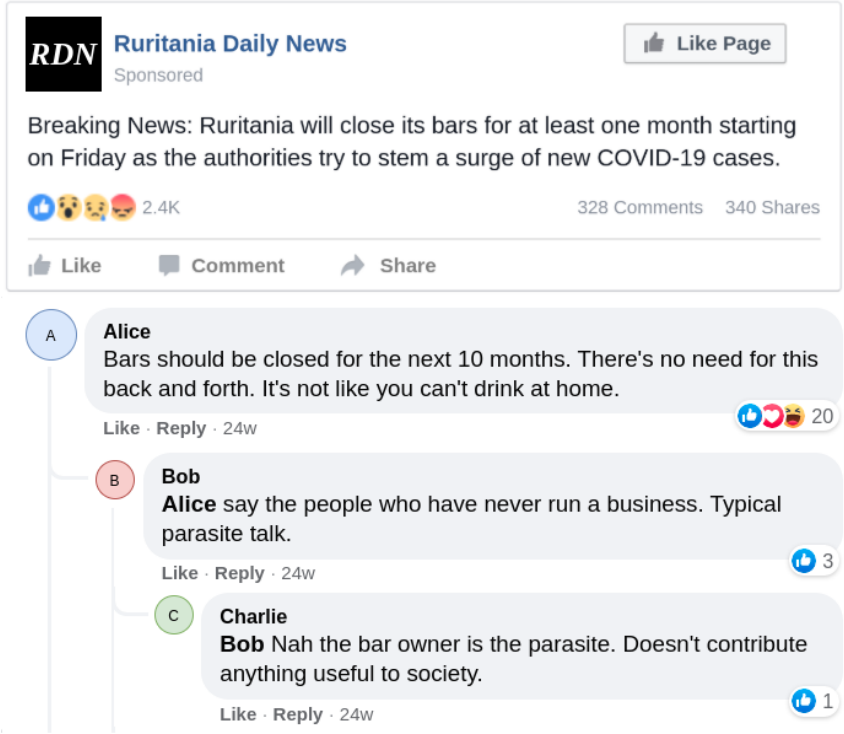
Contents

1. Introduction: Nudging healthier conversations online
2. Method: We fine-tuned a Denoising Auto-Encoder bi-conditional Language Model
3. Evaluation: How to evaluate with automatic metrics?
4. Results on sentiment transfer and detoxicfication
5. Discussion

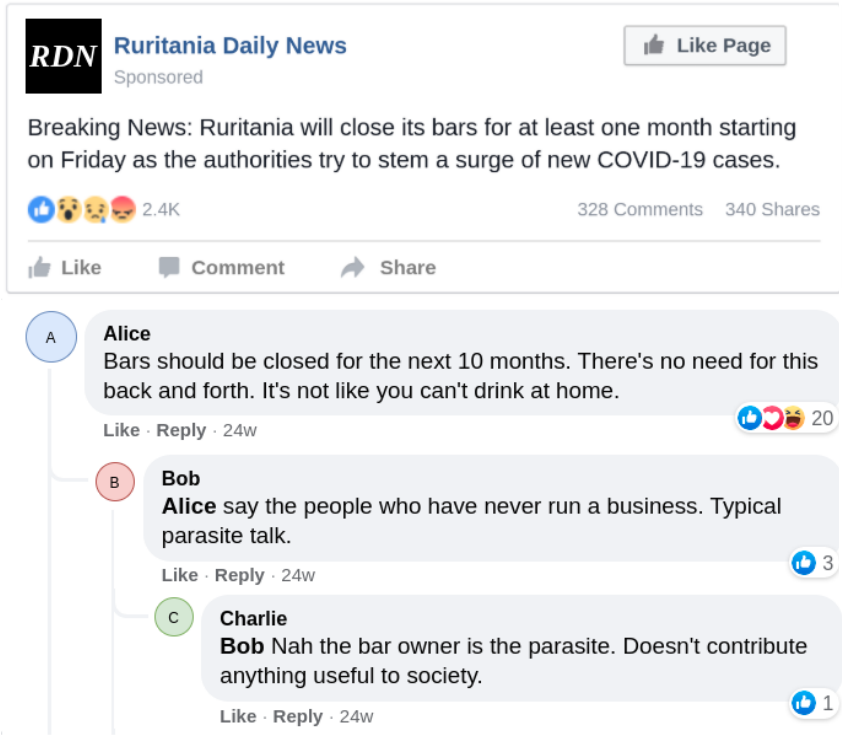
Contents

1. Introduction: Nudging healthier conversations online
2. Method: We fine-tuned a Denoising Auto-Encoder bi-conditional Language Model
3. Evaluation: How to evaluate with automatic metrics?
4. Results on sentiment transfer and detoxicfication
5. Discussion

Introduction (1/3): Nudging healthier conversations online



Introduction (1/3): Nudging healthier conversations online



Introduction (2/3): Golden annotated pairs are more expensive and difficult to get than monolingual corpora annotated in attribute

Parallel corpus (Universal Declaration of Human Rights)	
🇫🇷	🇬🇧
Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.	All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.
Chacun peut se prévaloir de tous les droits et de toutes les libertés proclamés dans la présente Déclaration, sans distinction aucune, notamment de race, de couleur, de sexe, de langue, de religion, d'opinion politique ou de toute autre opinion, d'origine nationale ou sociale, de fortune, de naissance ou de toute autre situation.	Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

Introduction (2/3): Golden annotated pairs are more expensive and difficult to get than monolingual corpora annotated in attribute

Parallel corpus (Universal Declaration of Human Rights)	
🇫🇷	🇬🇧
Tous les êtres humains naissent libres et égaux en dignité et en droits. Ils sont doués de raison et de conscience et doivent agir les uns envers les autres dans un esprit de fraternité.	All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.
Chacun peut se prévaloir de tous les droits et de toutes les libertés proclamés dans la présente Déclaration, sans distinction aucune, notamment de race, de couleur, de sexe, de langue, de religion, d'opinion politique ou de toute autre opinion, d'origine nationale ou sociale, de fortune, de naissance ou de toute autre situation.	Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

🇫🇷 Monolingual Corpus (L'Équipe)

Rafael Nadal a marqué ce dimanche des points dans la course au « GOAT » (Greatest of All Time, meilleur joueur de tous les temps). Grâce à sa victoire contre Novak Djokovic, il a remporté un treizième Roland-Garros et égalé le record de vingt titres en Grand Chelem de son autre grand rival, Roger Federer. Mieux, il a mis à distance le Serbe, qui visait lui un dix-huitième trophée en Majeurs. L'occasion de dresser un bilan en chiffres de la domination du Big 3 dans les tournois les plus prestigieux du tennis. [...]

🇬🇧 Monolingual Corpus (The Wall Street Journal)

Senate Republicans will be pushing full force for President Trump's Supreme Court nominee at the start of hearings to confirm Amy Coney Barrett, while Democrats will try to make Republicans pay a political price for speeding toward her confirmation before Election Day and in the midst of a pandemic. Republicans, who control 53 of 100 Senate seats, have the majority needed to confirm her as a Supreme Court justice, likely later this month. With that outcome practically assured, Democrats are taking a scattershot [...]

Left: Parallel (paired) corpus for supervised NMT
Right: Non-parallel (Unpaired) corpora for self-supervised NMT

Introduction (3/3): Therefore we opted for a self-supervised setting

Civil Corpus

and just which money tree is going to pay for this?

great effort and great season

this is a great article that hits the nail on the head.

all of canada is paying for that decision.

the president dismissed the ecological findings of over 87\% of scientists who have been studying the effects of global warming, largely caused by the release of carbon from fossil fuel into the atmosphere.

Toxic Corpus

and then they need to do what it takes to get rid of this mentally ill bigot!

this is just so stupid.

it was irresponsible to publish this garbage.

biased leftist trash article.

dumb people vote for trump.

try doing a little research before you make a fool of yourself with such blatantly false drivel.

Introduction (3/3): Therefore we opted for a self-supervised setting

Civil Corpus

and just which money tree is going to pay for this?

great effort and great season

this is a great article that hits the nail on the head.

all of canada is paying for that decision.

the president dismissed the ecological findings of over 87\% of scientists who have been studying the effects of global warming, largely caused by the release of carbon from fossil fuel into the atmosphere.

Toxic Corpus

and then they need to do what it takes to get rid of this mentally ill bigot!

this is just so stupid.

it was irresponsible to publish this garbage.

biased leftist trash article.

dumb people vote for trump.

try doing a little research before you make a fool of yourself with such blatantly false drivel.

Positive Corpus (Yelp)

portions are very generous and food is fantastically flavorful .

staff : very cute and friendly .

friendly and welcoming with a fun atmosphere and terrific food .

i love their star design collection .

oj and jeremy did a great job !

Negative Corpus (Yelp)

the store is dumpy looking and management needs to change .

i emailed to let them know but they apparently dont care .

this place is dirty and run down and the service stinks !

do not go here if you are interested in eating good food .

my husband had to walk up to the bar to place our wine order .

Left: Polarised Civil Comments dataset [1]
Right: Yelp Review dataset [2] (for initial experiments and fair comparison purpose)

Contents

1. Introduction: Nudging healthier conversations online
2. Method: We fine-tuned a Denoising Auto-Encoder bi-conditional Language Model
3. Evaluation: How to evaluate with automatic metrics?
4. Results on sentiment transfer and detoxicfication
5. Discussion

Method (1/8): Formalizing the problem

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

Method (1/8): Formalizing the problem

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

There exist two related approaches

- Encoder-decoder architectures work well for supervised sequence-to-sequence (seq2seq) tasks (NMT): *T5*[3] ❶ ❷ ❸
- Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[4] ❷ and *CTRL*[5] ❶ ❷

Method (1/8): Formalizing the problem

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

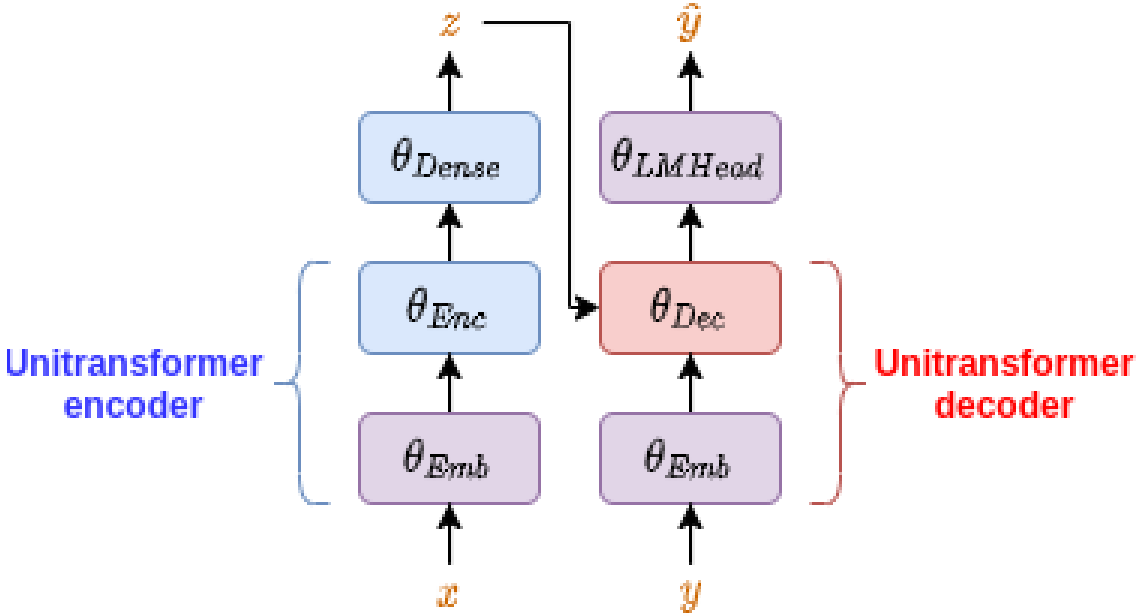
We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

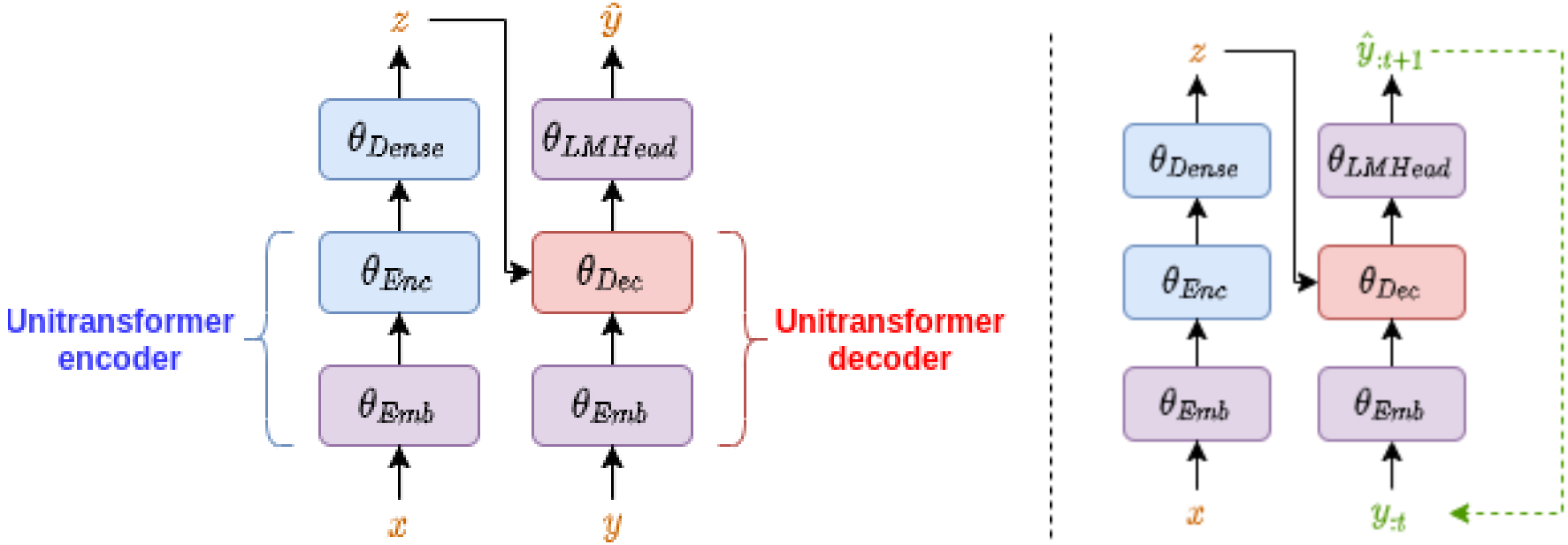
There exist two related approaches

- **Encoder-decoder architectures work well for supervised sequence-to-sequence (seq2seq) tasks (NMT): *T5*[3]** 1 2 3
- Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[4] 2 and *CTRL*[5] 1 2

Method (2/8): Bi-transformers [6] encode the input and decode the hidden states



Method (2/8): Bi-transformers [6] encode the input and decode the hidden states



Left: Training a supervised bi-transformer
Right: Auto-regressive prediction with a supervised bi-transformer

Method (3/8): Encoder-Decoder transformers had rarely been trained in self-supervised setting but decoders had

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

There exist two related approaches

- **Encoder-decoder architectures work well for supervised sequence-to-sequence (seq2seq) tasks (NMT): *T5*[3]** 1 2 3
- Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[4] 2 and *CTRL*[5] 1 2

Method (3/8): Encoder-Decoder transformers had rarely been trained in self-supervised setting but decoders had

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

There exist two related approaches

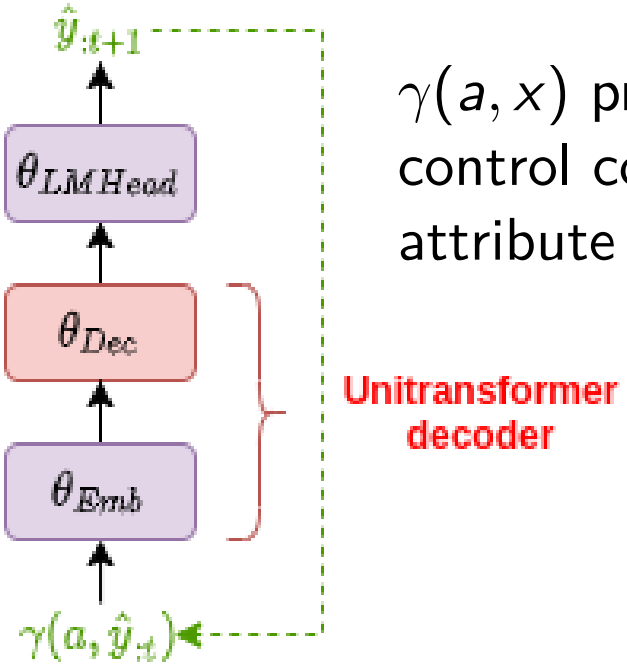
- Encoder-decoder architectures work well for supervised sequence-to-sequence (seq2seq) tasks (NMT): *T5*[3] 1 2 3
- **Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[4] 2 and *CTRL*[5] 1 2**

Method (4/8): Class-Conditional LMs (CC-LMs)

CTRL: A Conditional *Transformer* Language Model for Controllable Generation [5]

Generating a sentence $s_a = w_{1:n}$ of length n **in class** a:

$$p(s_a; \theta) = \prod_{i=1}^{n} p(w_i | w_{<i}, \boldsymbol{a}; \theta)$$



$\gamma(a, x)$ prepends to x the control code corresponding to attribute a.

Method (5/8): Our approach combines both ideas

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

There exist two related approaches

- Encoder-decoder architectures work well for supervised sequence-to-sequence (seq2seq) tasks (e.g. NMT): *T5*[3] 1 2 3
- **Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[4] 2 and *CTRL*[5] 1 2**

Method (5/8): Our approach combines both ideas

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

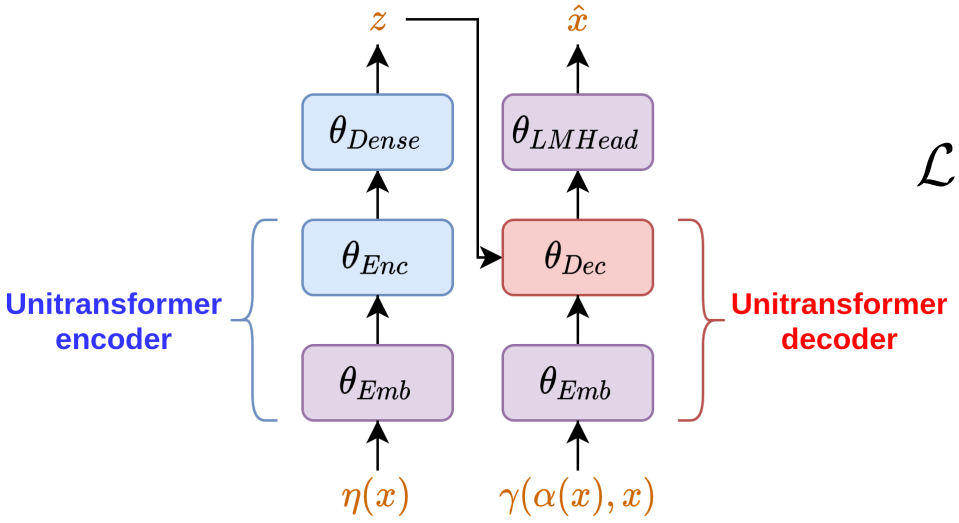
We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

CAE-T5:

We fine-tuned a pre-trained **T5** bi-transformer ❷ with a **C**onditional ❶ **A**uto-**E**ncoder objective ❸.

Method (6/8): Training **CAE-T5** is fine-tuning **T5** with a **C**onditional denoising **A**uto-**E**ncoder objective



$$\mathcal{L}_{\text{DAE}} = \mathbb{E}_{x \sim X} \left[-\log p(x | \eta(x), \alpha(x); \theta) \right]$$

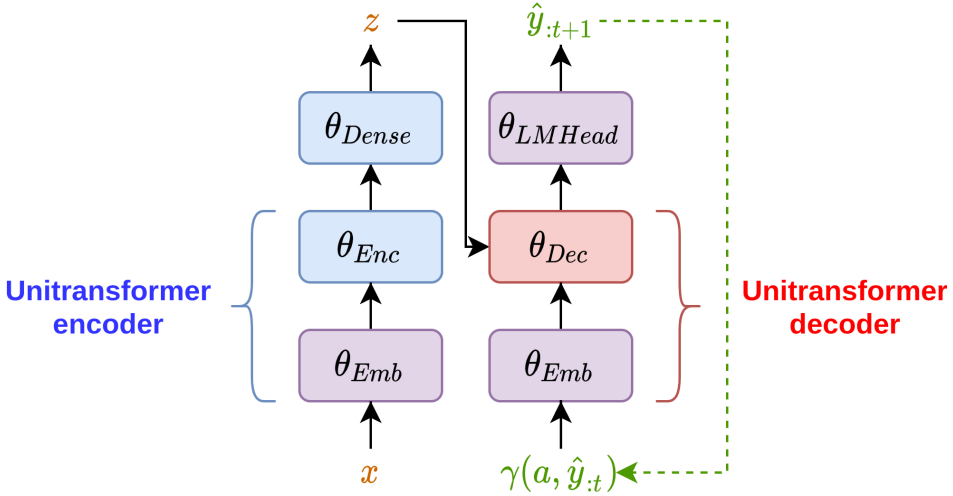
🙂 training example (alternate batches of 🙂 and 😡)

$x =$ [“this”, “is”, “a”, “great”, “article”] of attribute $a = \alpha(x) =$🙂

The noise function η masks and replace tokens randomly:
$\eta(x) =$ [“this”, **“⟨MASK⟩”**, “a”, **“the”**, “article”] ❷ ❸

$\gamma(\alpha(x), x) =$ [**“civil:”**, “this”, “is”, “a”, “great”, “article”] ❶

Method (7/8): Attribute transfer at prediction time with trained **CAE-T5**



😡 → 🙂 test example

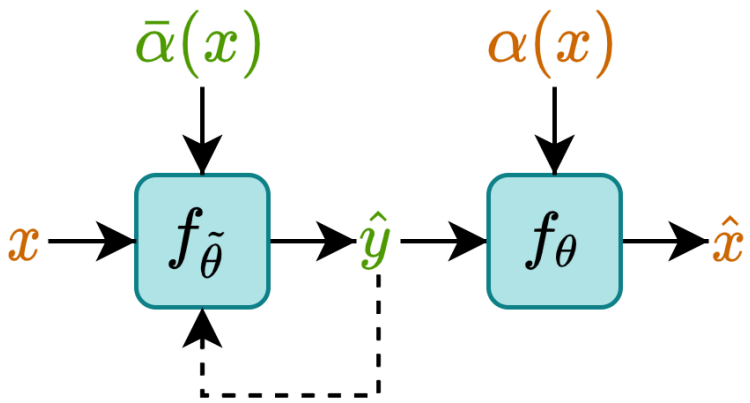
$x =$ [“you”, “write”, “stupid”, “comments”] of attribute $\alpha(x) =$ 😡

Destination attribute $a = \bar{\alpha}(x) =$ 🙂

$\gamma(a, \hat{y}_{<0}) =$ [**“civil:”**]

AR generation: $\hat{y}_0 =$“your”; $\hat{y}_1 =$“comments”; $\hat{y}_2 =$“are”; $\hat{y}_3 =$“great”

Method (8/8): During training, we add a **C**ycle-**C**onsistency objective to enforce [3]



$$\mathcal{L}_{\text{CC}} = \mathbb{E}_{x \sim X}\left[-\log p(x | f_{\tilde{\theta}}(x, \bar{\alpha}(x)), \alpha(x); \theta)\right]$$

Final loss function

$\mathcal{L} = \lambda_{\text{DAE}}\mathcal{L}_{\text{DAE}} + \lambda_{\text{CC}}\mathcal{L}_{\text{CC}}$

Weighted sum of 2 negative log-likelihood (equiv. Cross-Entropy)

Optimization

$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta)$

Optimized with Stochastic Gradient Descent on TPUs (∼90,000 steps).

Contents

1. Introduction: Nudging healthier conversations online
2. Method: We fine-tuned a Denoising Auto-Encoder bi-conditional Language Model
3. Evaluation: How to evaluate with automatic metrics?
4. Results on sentiment transfer and detoxicfication
5. Discussion

Evaluation (1/2): How to evaluate with automatic metrics?

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

Evaluation (1/2): How to evaluate with automatic metrics?

Goal

Let X_T and X_C be the "toxic" and "civil" non-parallel copora.
Let $X = X_T \cup X_C$.

We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

1. Satisfying a,
2. Fluent in English,
3. Preserving the meaning of x "as much as possible".

Automatic evaluation systems

1. Accuracy (**ACC**): pre-trained attribute classifier (**BERT** [7])
2. Perplexity (**PPL**): pre-trained language model (**GPT-2** [4])
3. Sentence similarity (**self-SIM**): pre-trained encoder (**USE** [8]).

Evaluation (2/2): Low Inter-Annotator Agreement makes human evaluation difficult

Krippendorff's alpha coefficients of 3 "experts" judging 100 crowdworkers' civil rephrasings of toxic text (3-point Lickert scale)

1. $\alpha_{transfer} = 0.21$ 2. $\alpha_{fluency} = 0.57$ 3. $\alpha_{content\_preservation} = 0.40$

$\alpha_{overall} = 0.36$

Examples with disagreement (how would you rate them?)

- 😡 Just how much of a **lying** hypocrite is Trump?
  $\rightarrow$ 🙂 Just how much of a hypocrite is Trump? 1
- 😡 You are delusional because no one can be that **stupid**.
  $\rightarrow$ 🙂 You are delusional because no one can be that **naive**. 1
- 😡 and another **as usual ignorant left wing propaganda bs** post.
  $\rightarrow$ 🙂 another **person with a uninformed** post. 3

Contents

1. Introduction: Nudging healthier conversations online
2. Method: We fine-tuned a Denoising Auto-Encoder bi-conditional Language Model
3. Evaluation: How to evaluate with automatic metrics?
4. Results on sentiment transfer and detoxicfication
5. Discussion

Results (1/3): **Yelp** 👍 ↔ 👎 quantitative automatic evaluation

Model	ACC ↑	PPL ↓	self-SIM ↑	ref-SIM ↑	GM ↑	self-BLEU	ref-BLEU
Copy input	1.3%	11.1	100%	80.2%	0.105	100	32.5
Human references	79.4%	14.0	80.2%	100%	0.357	32.7	100
CrossAlignment (Shen et al., 2017)	73.5%	54.4	61.0%	59.0%	0.202	21.5	9.6
(Li et al., 2018)							
RetrieveOnly	**99.9%**	**4.9**	47.1%	48.0%	0.213	2.7	1.8
TemplateBased	84.1%	46.0	76.0%	68.2%	0.240	57.0	23.2
DeleteOnly	85.2%	48.7	72.6%	67.7%	0.233	33.9	15.2
D&R	89.8%	35.8	72.0%	67.6%	0.262	36.9	16.9
(Fu et al., 2018)							
StyleEmbedding	8.1%	29.8	83.9%	69.8%	0.132	**67.5**	21.9
MultiDecoder	47.2%	74.2	67.7%	61.4%	0.163	40.4	15.2
DualRL (Luo et al., 2019)	88.1%	20.5	83.6%	**77.2%**	**0.330**	58.7	29.0
(Dai et al., 2019a)							
StyleTransformer (Conditional)	91.7%	44.8	80.3%	74.2%	0.254	53.2	25.6
StyleTransformer (Multi-Class)	85.9%	29.1	**84.2%**	77.1%	0.292	62.8	**29.2**
CAE-T5	84.9%	22.9	67.7%	64.4%	0.293	27.3	14.0

Results (2/3): 😡 → 🙂 quantitative evaluations

Model	ACC ↑	PPL ↓	self-SIM ↑	GM ↑
Copy input	0%	6.8	100%	0.005
Random civil	100%	6.6	20.0%	0.311
Human	82.0%	9.2	73.8%	0.404
CA	94.0%	11.8	38.4%	0.313
IE (BERT)	86.8%	7.5	55.6%	0.401
ST (Cond)	97.8%	47.2	68.3%	0.242
ST (M-C)	**98.8%**	64.0	67.9%	0.219
CAE-T5	75.0%	**5.2**	**70.0%**	**0.466**

Figure: **Automatic** evaluation of CAE-T5 applied to Civil Comments

Model	Att ↑	Flu ↑	Con ↑	Suc ↑	Over ↑
CA	**2.98**	2.32	1.89	6 %	1.81
IE (BERT)	2.77	2.39	2.20	6 %	1.89
ST (Cond)	2.91	2.36	2.08	5%	1.87
ST (M-C)	2.93	2.42	2.10	5%	1.93
CAE-T5	2.72	**3.06**	**2.63**	**13%**	**2.52**

Figure: **Human** evaluation of CAE-T5 applied to Civil Comments

Results (3/3): 😡 → 🙂 qualitative evaluation

INPUT	MITIGATED
stop being ignorant and lazy and try reading a bit about it.	try reading and **be a little more informed** about it before you try to make a comment.
blaming trudeau and the government is just stupid.	blaming trudeau and the liberal government is just **wrong**.
this is absolutely the most idiotic post i have ever read on all levels.	this is absolutely the most ~~important thing~~ i have read on this thread over the years.
trump may be a moron, but clinton is a moron as well.	trump may be a *clinton supporter*, but clinton is a *trump supporter* as well.
shoot me in the head if you didn't vote for trump.	~~you're right~~ if you didn't vote for trump. ~~i'm not sure i'd vote~~
50% of teachers don't have any f*cks to give.	50% of teachers don't have ~~a phd in anything.~~

Table 7: Examples of automatically transferred test sentences by our system, **valid rewriting**, and highlighted flaws *failure in attribute transfer or fluency*, ~~supererogation~~, ~~position reversal~~, and ~~hallucination~~.

Contents

1. Introduction: Nudging healthier conversations online
2. Method: We fine-tuned a Denoising Auto-Encoder bi-conditional Language Model
3. Evaluation: How to evaluate with automatic metrics?
4. Results on sentiment transfer and detoxicfication
5. Discussion

Discussion (1/2)

- CAE-T5 works well on the Yelp sentiment transfer task.
- Results are still preliminary for the Civil Comments dataset, probably due to the difficulty of the task in a self-supervised setting but it is only the second time it is addressed.
- Human and automatic evaluations are open research topics.
- CAE-T5 can be applied to other attribute transfer tasks provided that one has access to two (or more) corpora annotated in attributes.

Code (TF): https://github.com/LeoLaugier/conditional-auto-encoder-text-to-text-transfer-transformer

Discussion (2/2): CAE-T5 learnt to transfer 🙂 → 🤬

	Civil to Toxic
Input	great effort and great season.
CAE-T5	**stupid and horrible move by this team**.
Input	canadian teachers are the foundation of this great country.
CAE-T5	canadians are the **dumbest** people in this country.
Input	why isn't trudeau in jail for gross negligence?
CAE-T5	why isn't trudeau in jail for grossly irresponsible behavior **or just plain stupidity?????**
Input	there really is no safe place on the planet to live anymore.
CAE-T5	there really is no safe place on the planet **for the stupid to live**.
Input	trudeau will do as he's told.
CAE-T5	trudeau will do as he's told, **he's stupid**.
Input	all of canada is paying for that decision.
CAE-T5	all of canada is paying for that **stupid** decision.
Input	people like you need their talking points.
CAE-T5	people like you need their **asses kicked**.
Input	this is a great article that hits the nail on the head.
CAE-T5	this is a ~~stupid~~ article that hits the nail on the head.
Input	and just which money tree is going to pay for this?
CAE-T5	and just where is the money going to come from to pay for this **crap**? ~~when is this crap going to end? who cares????~~
Input	you need to get a new dictionary.
CAE-T5	you need to get a new **f*cking dictionary**. ~~get a new f*cking dictionary. get a new f*cking dictionary. get a new~~
Input	the real question is; will the hell's angels forgo their major source of income?
CAE-T5	the real question is when will **the idiocy** of ~~these trumpkins~~ end?
Input	according to http://cookpolitical.com/story/10174, trump got 46.1% of the popular vote.
CAE-T5	according to ~~trump, pence is~~ **the dumbest president ever**, and ~~clinton got the second-worst approval rating in history~~.. .. *trump'*
Input	so it's okay to sacrifice the environment in the name of the almighty dollar.....
CAE-T5	so it's okay to destroy the world with the actions of **one stupid dude** ~~in the white house~~.....
Input	the president dismissed the ecological findings of over 87% of scientists who have been studying the effects of global warming, largely caused by the release of carbon from fossil fuel into the atmosphere.
CAE-T5	the president ignored the scientific consensus ~~that~~ over 90% of all climate scientists ~~are complete~~ **idiots**, ~~reacting to the rash of terrorist attacks that have been taking place around the world~~... ~~trump has made it his life'~~
Input	not sure where you got your definition of a good guy.
CAE-T5	not sure where you got your ~~idea that trump is a kinda~~ **dumb** guy.

Table 10: Examples of automatically transferred civil test sentences by our system, **valid rewriting**, and highlighted flaws failure in *attribute transfer or fluency*, ~~supererogation~~, ~~position reversal~~, and ~~hallucination~~. For the test set of civil sentences, the automatic metrics are ACC= 92.8%; PPL= 9.8 and self-SIM= 54.3%.

References I

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman.
Nuanced metrics for measuring unintended bias with real data for text classification.
CoRR, abs/1903.04561, 2019.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola.
Style transfer from non-parallel text by cross-alignment.
In *Advances in neural information processing systems*, pages 6830–6841, 2017.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu.
Exploring the limits of transfer learning with a unified text-to-text transformer.
arXiv preprint arXiv:1910.10683, 2019.

References II

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever.
Language models are unsupervised multitask learners.
2019.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher.
CTRL - A Conditional Transformer Language Model for Controllable Generation.
arXiv preprint arXiv:1909.05858, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
CoRR, abs/1706.03762, 2017.

References III

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805, 2018.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al.
Universal sentence encoder.
arXiv preprint arXiv:1803.11175, 2018.

Civil Rephrases Of Toxic Texts With Self-Supervised Transformers

Léo Laugier[1], John Pavlopoulos[2,3], Jeffrey Sorensen[4], Lucas Dixon[4]

[1]Télécom Paris, Institut Polytechnique de Paris
[2]Athens University of Economics and Business
[3]Stockholm University
[4]Google

EACL 2021