DIG Seminar: Civil Rephrases Of Toxic Texts With Self-Supervised Transformers

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

[1]Télécom Paris, Institut Polytechnique de Paris

[2]Athens University of Economics & Business

[3]Stockholm University

[4]Google

October 15, 2020

Contents

Contents

Introduction (1/5): Nudging healthier conversations online

The New York Times
October 5 at 2:51 PM ·

All bars in Paris will close for at least two weeks starting on Tuesday as the authorities try to stem a surge of new cases in the French capital.

NYTIMES.COM
Paris will close its bars for at least two weeks starting on Tuesday.

2.4K   328 Comments 340 Shares

Like   Comment   Share

All Comments

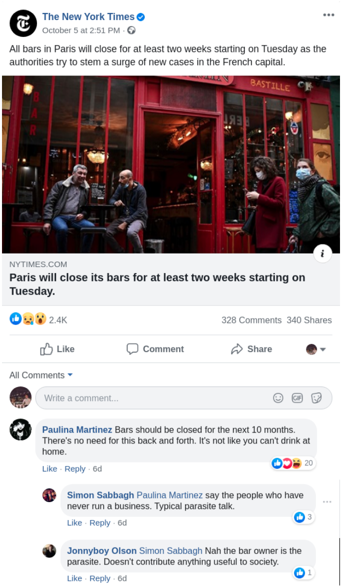
Write a comment...

Paulina Martinez Bars should be closed for the next 10 months. There's no need for this back and forth. It's not like you can't drink at home.
20
Like · Reply · 6d

Simon Sabbagh Paulina Martinez say the people who have never run a business. Typical parasite talk.
3
Like · Reply · 6d

Jonnyboy Olson Simon Sabbagh Nah the bar owner is the parasite. Doesn't contribute anything useful to society.
1
Like · Reply · 6d

Introduction (1/5): Nudging healthier conversations online



Your comment could be rephrased in a more civil manner: "@Paulina Martinez besides customers, I think you should consider that business owners struggle."

Introduction (2/5): Machine learning systems classify toxic comments online

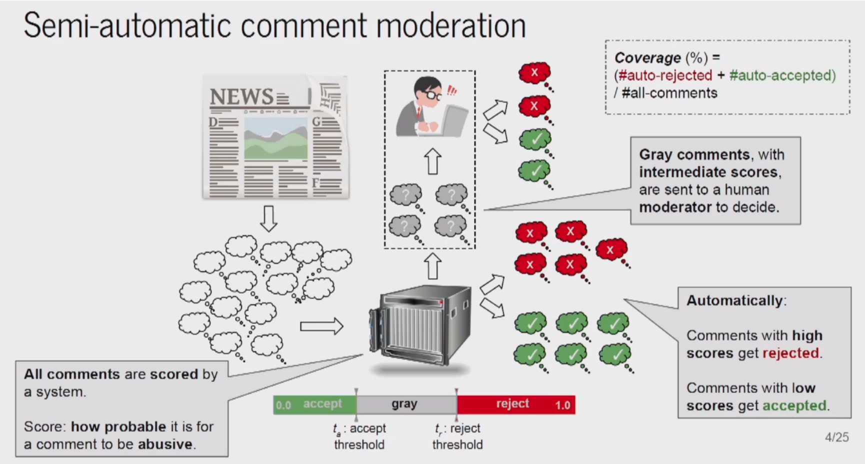
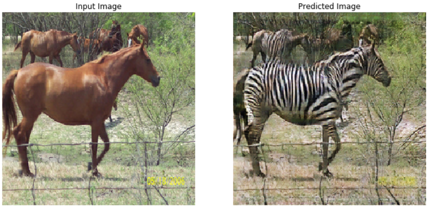


Figure: from Pavlopoulos et al. [1]

Introduction (3/5): Deep learning is efficient when applied to generative transfer tasks



Introduction (3/5): Deep learning is efficient when applied to generative transfer tasks

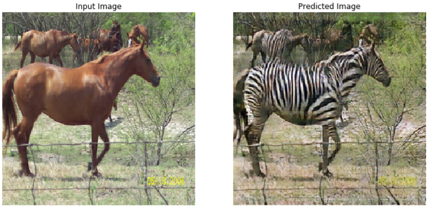


Figure:
Left: CycleGAN [2]
Right: Neural Machine Translation (NMT) (from https://jalammar.github.io/)

Introduction (4/5): 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 (4/5): 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 [...]

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

Introduction (5/5): 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 (5/5): Therefore we opted for a self-supervised setting

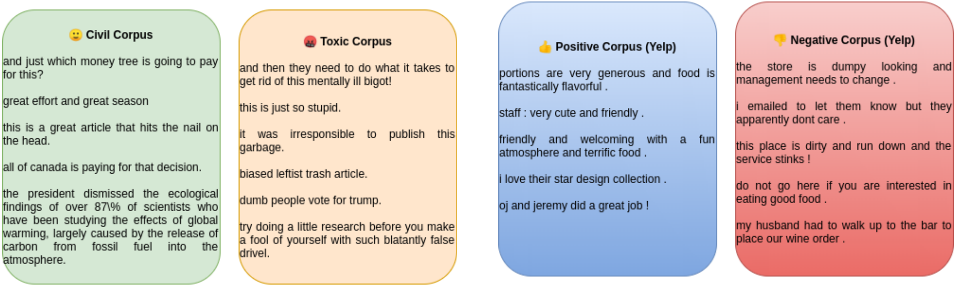


Figure:
Left: Polarised Civil Comments dataset [3]
Right: Yelp Review dataset [4] (for initial experiments and fair comparison purpose)

Contents

Method (1/14): 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 (1/14): 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*[5] 1 2 3
- Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[6] 2 and *CTRL*[7] 1 2

Method (1/14): Formalizing the problem

Goal

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

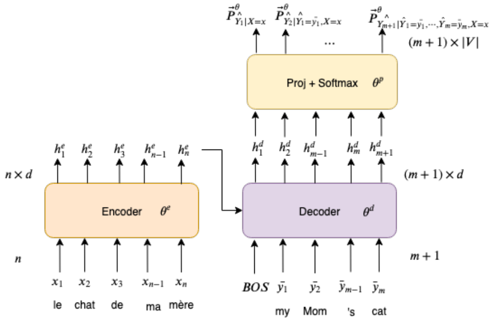
We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

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

There exist two related approaches

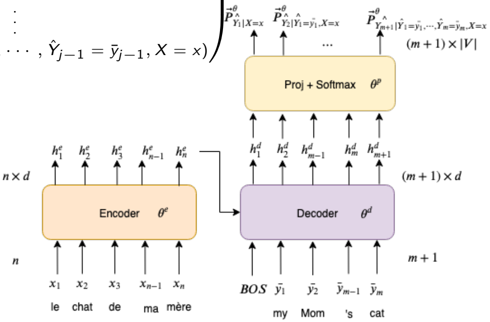
- **Encoder-decoder architectures work well for supervised sequence-to-sequence (seq2seq) tasks (NMT):** ***T5*[5]** 1 2 3
- Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[6] 2 and *CTRL*[7] 1 2

$$\bar{y}_j = \left\{ \; y_j \text{ if training} \right.$$



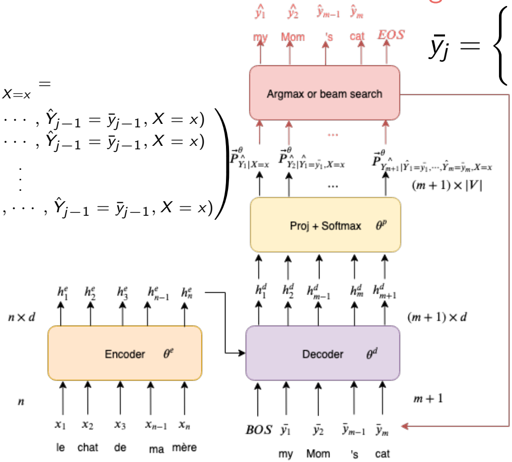
$$\bar{y}_j = \left\{ \; y_j \text{ if training} \right.$$

$$\vec{P}^{\theta}_{\hat{Y}_j|\hat{Y}_1=\bar{y}_1,\cdots,\hat{Y}_{j-1}=\bar{y}_{j-1},X=x} = \begin{pmatrix} P^{\theta}(\hat{Y}_j = w_1 | \hat{Y}_1 = \bar{y}_1, \cdots, \hat{Y}_{j-1} = \bar{y}_{j-1}, X = x) \\ P^{\theta}(\hat{Y}_j = w_2 | \hat{Y}_1 = \bar{y}_1, \cdots, \hat{Y}_{j-1} = \bar{y}_{j-1}, X = x) \\ \vdots \\ P^{\theta}(\hat{Y}_j = w_{|V|} | \hat{Y}_1 = \bar{y}_1, \cdots, \hat{Y}_{j-1} = \bar{y}_{j-1}, X = x) \end{pmatrix}$$



Method (2/14): Encoder-Decoder for **supervised** seq2seq

$$\vec{P}^{\theta}_{\hat{Y}_j|\hat{Y}_1=\bar{y}_1,\cdots,\hat{Y}_{j-1}=\bar{y}_{j-1},X=x} = \begin{pmatrix} P^{\theta}(\hat{Y}_j = w_1 | \hat{Y}_1 = \bar{y}_1, \cdots, \hat{Y}_{j-1} = \bar{y}_{j-1}, X = x) \\ P^{\theta}(\hat{Y}_j = w_2 | \hat{Y}_1 = \bar{y}_1, \cdots, \hat{Y}_{j-1} = \bar{y}_{j-1}, X = x) \\ \vdots \\ P^{\theta}(\hat{Y}_j = w_{|V|} | \hat{Y}_1 = \bar{y}_1, \cdots, \hat{Y}_{j-1} = \bar{y}_{j-1}, X = x) \end{pmatrix}$$



Auto-Regressive (AR) generation at inference

$$\bar{y}_j = \begin{cases} y_j \text{ if training} \\ \hat{y}_j \text{ if inference} \end{cases}$$

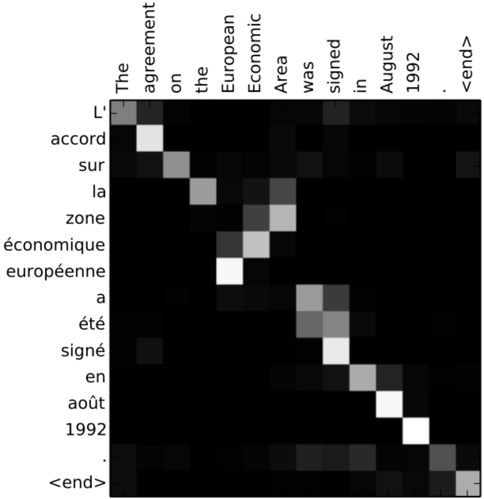


Figure: **Cross-attention** heat map for NMT, from Bahdanau *et al.* [8] (2015)

Method (3/14): Encoding and decoding is modeled *via* **attention** mechanism (see https://jalammar.github.io/)

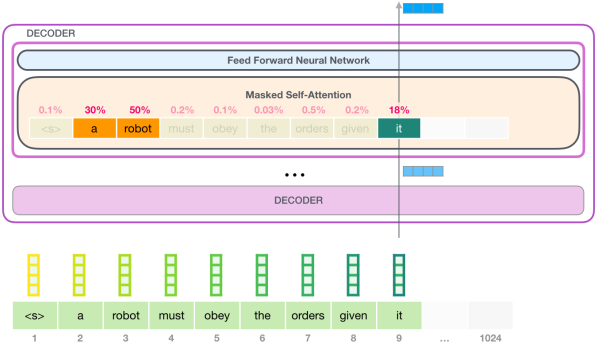
Second Law of Robotics

A robot must obey the orders given *it* by human beings except where *such orders* would conflict with *the First Law*.

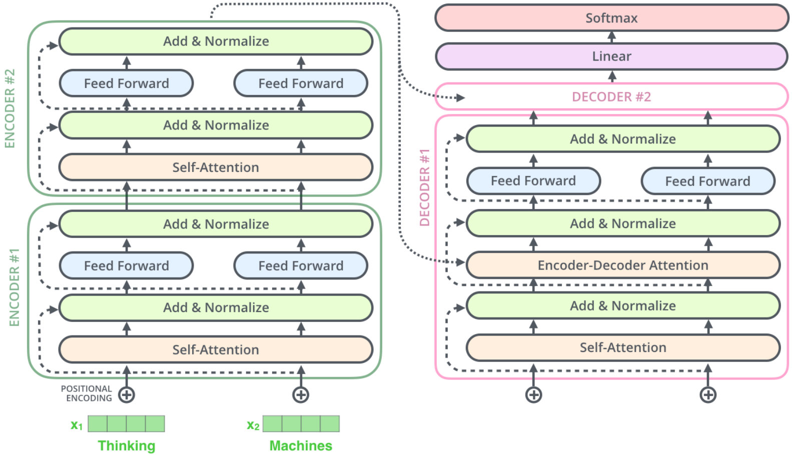
Method (3/14): Encoding and decoding is modeled *via* **attention** mechanism (see https://jalammar.github.io/)

Second Law of Robotics

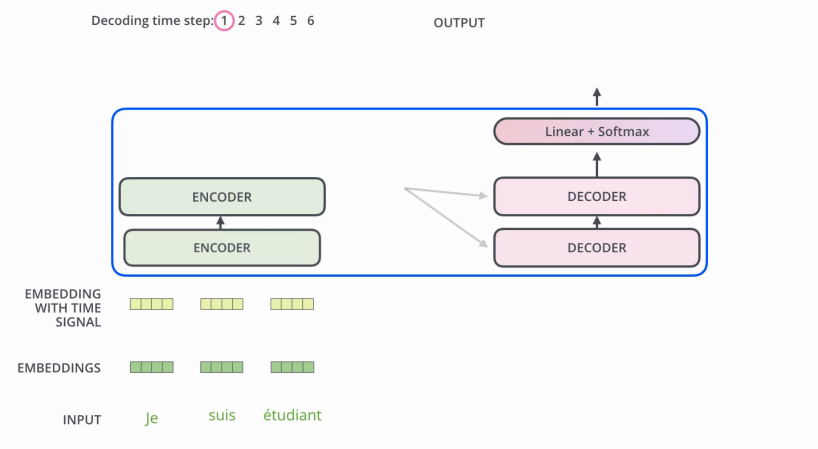
A robot must obey the orders given *it* by human beings except where *such orders* would conflict with *the First Law*.



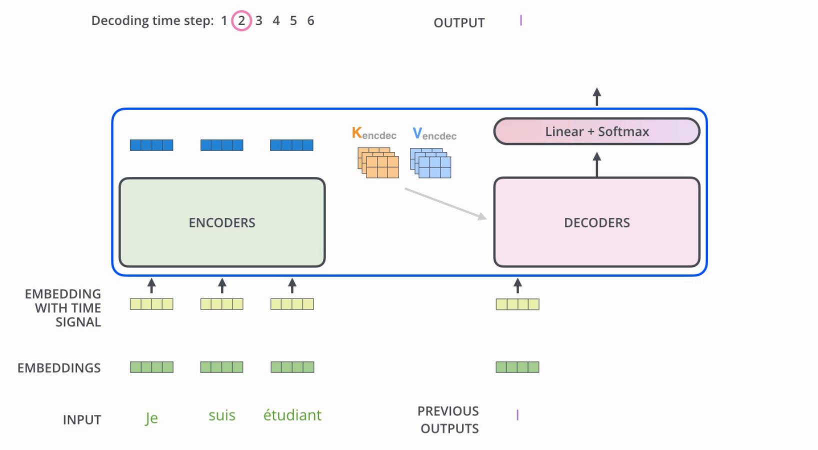
Method (4/14): Bi-transformers [9] encode the input and decode the hidden states (see https://jalammar.github.io/)



Method (5/14): Inference time - where the Natural Language Generation happens (see https://jalammar.github.io/)



Method (5/14): Inference time - where the Natural Language Generation happens (see https://jalammar.github.io/)



Method (6/14): Transformers learn relevant features

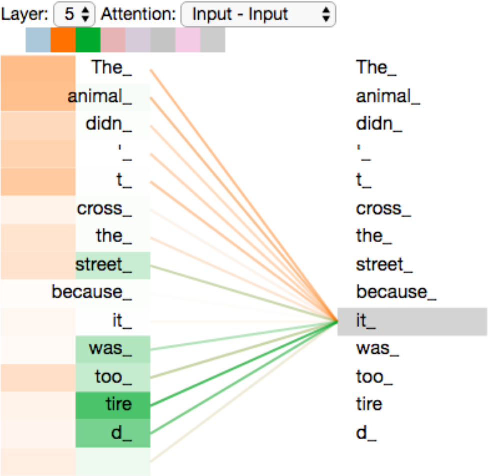


Figure: As we encode the word "it", one attention head is focusing most on "the animal", while another is focusing on "tired".

Method (7/14): Transformers benefit from **scaling** their size (hidden size and depth) and **pre-training** on massive corpus: *T5*[5]

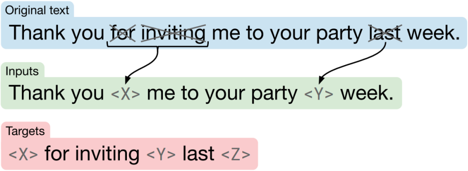


Figure: **Transfer learning**: **T**ext-to-**T**ext **T**ransfer **T**ransformer (**T5**) is pre-trained with a self-supervised objective to learn semantic representations, before being fine-tuned on downstream supervised tasks (NMT, sentiment analysis, etc.)

Pre-training dataset: "**C**olossal **C**lean **C**rawled **C**orpus" (**C4**) ∼34 Billion tokens (∼750 GB) of clean English text scraped from the web.
T5 sizes: Small, Base, **Large** (24 layers; 770 Million parameters), 3B, 11B.

Method (8/14): 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*[5]** 1 2 3
- Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[6] 2 and *CTRL*[7] 1 2

Method (8/14): 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*[5] 1 2 3
- **Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[6] 2 and *CTRL*[7] 1 2**

Method (9/14): Introduction to Language Models (LM)

What is a Language Model?

A statistical Language Model is a probability distribution over sequences of words.

Predicting the next word: $p(w_t|w_{<t})$
If $w_{<t} =$ [“the”,“best”,“place”,“to”,“visit”,“in”,“France”,“is”] then

$p(\text{“Paris”}|w_{<t}) = 0.6$
$p(\text{“Mont”}|w_{<t}) = 0.3$
$p(\text{“Saclay”}|w_{<t}) = \epsilon$
$p(\text{“have”}|w_{<t}) = 0$

Method (9/14): Introduction to Language Models (LM)

What is a Language Model?

A statistical Language Model is a probability distribution over sequences of words.

Predicting the next word: $p(w_t|w_{<t})$
If $w_{<t} =$ [“the”,“best”,“place”,“to”,“visit”,“in”,“France”,“is”] then

$p(\text{“Paris”}|w_{<t}) = 0.6$
$p(\text{“Mont”}|w_{<t}) = 0.3$
$p(\text{“Saclay”}|w_{<t}) = \epsilon$
$p(\text{“have”}|w_{<t}) = 0$

Deep learning provides parametric architectures able to learn in a **self-supervised** setting to approximate LMs: $p(w_t|w_{<t};\boldsymbol{\theta})$. They are trained with **maximum likelihood** on massive corpora like C4.

Generating $w_{\geq t}$ from prompt $w_{<t}$: $p(w_{\geq t}|w_{<t};\theta) = \prod_{i=t}^{n} p(w_i|w_{<i};\theta)$

Method (10/14): Class-Conditional LMs (CC-LMs)

CTRL: A Conditional *Transformer* Language Model for Controllable Generation [7]

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

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

If the "**prompt**" $w_{<t} = [\text{"Paris"}, \text{"is"}]$ and $a \in \{$👍;👎$\}$ then

$$\arg\max_{w_{t:t+4}} p(w_{t:t+4} | w_{<t}, a = \text{👍}; \theta) = [\text{"such"}, \text{"a"}, \text{"beautiful"}, \text{"city"}]$$

$$\arg\max_{w_{t:t+4}} p(w_{t:t+4} | w_{<t}, a = \text{👎}; \theta) = [\text{"a"}, \text{"very"}, \text{"boring"}, \text{"town"}]$$

Method (11/14): 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*[5] 1 2 3
- **Language Models (LMs) are efficient for self-supervised "free" generation: *GPT-2*[6] 2 and *CTRL*[7] 1 2**

Method (11/14): Our approach combines both ideas

Goal

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

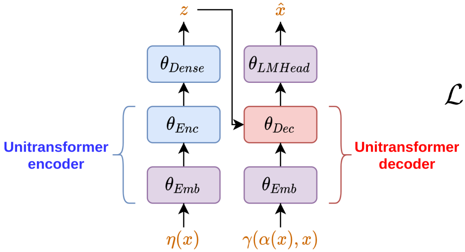
We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

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 2 with a **C**onditional 1 **A**uto-**E**ncoder objective 3.

Method (12/14): Training **CAE-T5** is fine-tuning **T5** with a **C**onditional denoising **A**uto-**E**ncoder objective



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

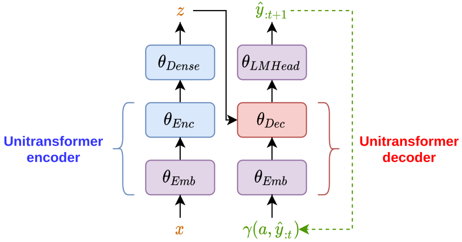
🙂 **training example** (alternate batches of 🙂 and 🤬)

$x =$ ["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

$\gamma(a, x)$ prepends to x the control code corresponding to attribute a:
$\gamma(\alpha(x), x) =$ [**"civil:"**, "this", "is", "a", "great", "article"] ❶

Method (13/14): Attribute transfer at prediction time with trained **CAE-T5**



🤬 $\rightarrow$ 🙂 test example

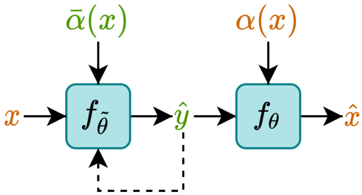
x = [“you”, “write”, “stupid”, “comments”] of attribute $\alpha(x) =$ 🤬

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 (14/14): During training, we add a Cycle-Consistency 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

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 $X = X_T \cup X_C$.

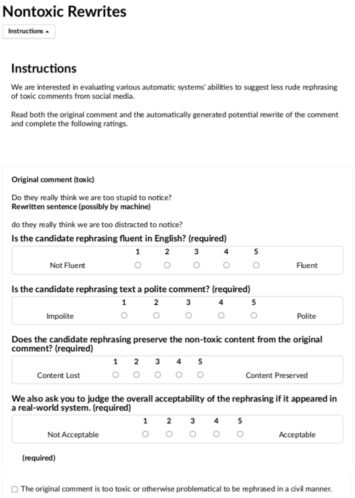
We aim at learning in a **self-supervised** setting, a mapping f_θ s. t. $\forall (x, a) \in X \times \{\text{"civil"}, \text{"toxic"}\}$, $y = f_\theta(x, a)$ is a text:

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** [10])
2. Perplexity (**PPL**): pre-trained language model (**GPT-2** [6])
3. Sentence similarity (**self-SIM**): pre-trained encoder (**USE** [11]).

Evaluation (2/2): Human evaluation through crowdworking



Nontoxic Rewrites

Instructions ▴

Instructions

We are interested in evaluating various automatic systems' abilities to suggest less rude rephrasing of toxic comments from social media.

Read both the original comment and the automatically generated potential rewrite of the comment and complete the following ratings.

Original comment (toxic)

Do they really think we are too stupid to notice?

Rewritten sentence (possibly by machine)

do they really think we are too distracted to notice?

Is the candidate rephrasing fluent in English? (required)

	1	2	3	4	5	
Not Fluent	○	○	○	○	○	Fluent

Is the candidate rephrasing text a polite comment? (required)

	1	2	3	4	5	
Impolite	○	○	○	○	○	Polite

Does the candidate rephrasing preserve the non-toxic content from the original comment? (required)

	1	2	3	4	5	
Content Lost	○	○	○	○	○	Content Preserved

We also ask you to judge the overall acceptability of the rephrasing if it appeared in a real-world system. (required)

	1	2	3	4	5	
Not Acceptable	○	○	○	○	○	Acceptable

(required)

☐ The original comment is too toxic or otherwise problematical to be rephrased in a civil manner.

Figure: Guidelines provided to annotators on Appen

Contents

Results (1/4): **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 (2/4): **Yelp** 👍 ↔ 👎 qualitative evaluation

	Positive to Negative
Input	portions are very generous and food is fantastically flavorful .
DualRL	portions are very **thin** and food is *confusing* .
ST (Multi)	portions are very *poorly* and food is *springs* **flavorless** .
CAE-T5	portions are very **small** and food is **awfully greasy for the price** .
Human	portions are very **small** and food is **not flavorful** .
Input	staff : very cute and friendly .
DualRL	staff : very **awful** and **rude** .
ST (Multi)	staff : very *nightmare* and *poor* .
CAE-T5	staff : very **rude** and **pushy** .
Human	staff : very **ugly** and **mean** .
Input	friendly and welcoming with a fun atmosphere and terrific food .
DualRL	**rude** and **unprofessional** with a **loud** atmosphere and **awful** food .
ST (Multi)	**poor** and **fake** with a *fun* atmosphere and **mushy** food .
CAE-T5	**rude** and **unhelpful service** with **a forced smile** and *attitude* .
Human	**unfriendly** and **unwelcoming** with a **bad** atmosphere and food .
Input	i love their star design collection .
DualRL	i **hate** their star design *disgrace* .
ST (Multi)	i *do n't care* star ~~bites~~ collection .
CAE-T5	i **hate** ~~starbucks corporate~~ . ~~the staff is horrible~~ .
Human	i **ca n't stand** their star design collection .
Input	oj and jeremy did a great job !
DualRL	oj and jeremy did a *great* job ! ~~disgrace ! disgrace~~ !
ST (Multi)	oj and jeremy did a **terrible** job !
CAE-T5	~~oj~~ and ~~jesus christ i did n't have any change~~ !
Human	oj and jeremy did a **terrible** job !
	Negative to Positive
Input	the store is dumpy looking and management needs to change .
DualRL	the store is **perfect** looking and management *speaks to change perfectly* .
ST (Multi)	the store is **dumpy** looking and management *moved to change* .
Ours	the store is **neatly organized and clean** and **staff is on top of it** .
Human	managment **is top notch** , the **place looks great** .
Input	i emailed to let them know but they apparently dont care .
DualRL	i *loved them know them know but they dont care* .
ST (Multi)	i emailed to let them know but they *honestly played their* .
CAE-T5	i emailed to let them know **and** they **happily responded right away . a great service**
Human	i emailed to let them know **they really do care** .
Input	this place is dirty and run down and the service stinks !
DualRL	this place is **clean** and run *perfect* and the service **helped** !
ST (Multi)	this place is *quick* and *run down* and the service *stunning* !
CAE-T5	this place is **clean** and **well maintained** and the service **is great ! ! !**
Human	this place is **clean** , **not run down** , and the service **was great** .
Input	do not go here if you are interested in eating good food .
DualRL	**definitely go here** if you are interested in eating good food .
ST (Multi)	*do not go here* if you are interested in eating good food .
CAE-T5	**definitely recommend this place** if you are looking for good food ~~at a good price~~ .
Human	*do not go here* if you are interested in eating *bad* food .
Input	my husband had to walk up to the bar to place our wine order .
DualRL	my husband *had to* walk up to the bar to place our drink order *perfectly* .
ST (Multi)	my husband *had to* walk up to the bar to place our wine order .
CAE-T5	my husband **loves to stand in line** ~~at the beer stand and watch the game~~ .
Human	my husband **did n't have to** walk up to the bar to place our wine order .

Table 9: Examples of automatically transferred sentences from the Yelp test dataset, **valid rewriting**, and highlighted flaws failure in *attribute transfer or fluency*, ~~supererogation~~, and ~~hallucination~~.

Results (3/4): 🤬 → 🙂 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

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.
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

Conclusion (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.

Currently under review at EACL 2021.
Code (TF): https://github.com/LeoLaugier/conditional-auto-encoder-text-to-text-transfer-transformer

	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

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos.
Deeper attention to abusive user content moderation.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros.
Unpaired image-to-image translation using cycle-consistent adversarial networks.
In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

References II

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.

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio.
Neural machine translation by jointly learning to align and translate.
In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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.

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.

References V

Yulia Tsvetkov.
Towards personalized adaptive nlp: Modeling output spaces in continuous-output language generation.
2019.

Yoon Kim.
Convolutional neural networks for sentence classification.
In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le.
Sequence to sequence learning with neural networks.
In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014.

References VI

Jeffrey Pennington, Richard Socher, and Christopher D. Manning.
Glove: Global vectors for word representation.
In *In EMNLP*, 2014.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov.
Transformer-xl: Attentive language models beyond a fixed-length context.
CoRR, abs/1901.02860, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le.
Xlnet: Generalized autoregressive pretraining for language understanding.
CoRR, abs/1906.08237, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
Roberta: A robustly optimized BERT pretraining approach.
CoRR, abs/1907.11692, 2019.

DIG Seminar: Civil Rephrases Of Toxic Texts With Self-Supervised Transformers

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

[1]Télécom Paris, Institut Polytechnique de Paris

[2]Athens University of Economics & Business

[3]Stockholm University

[4]Google

October 15, 2020