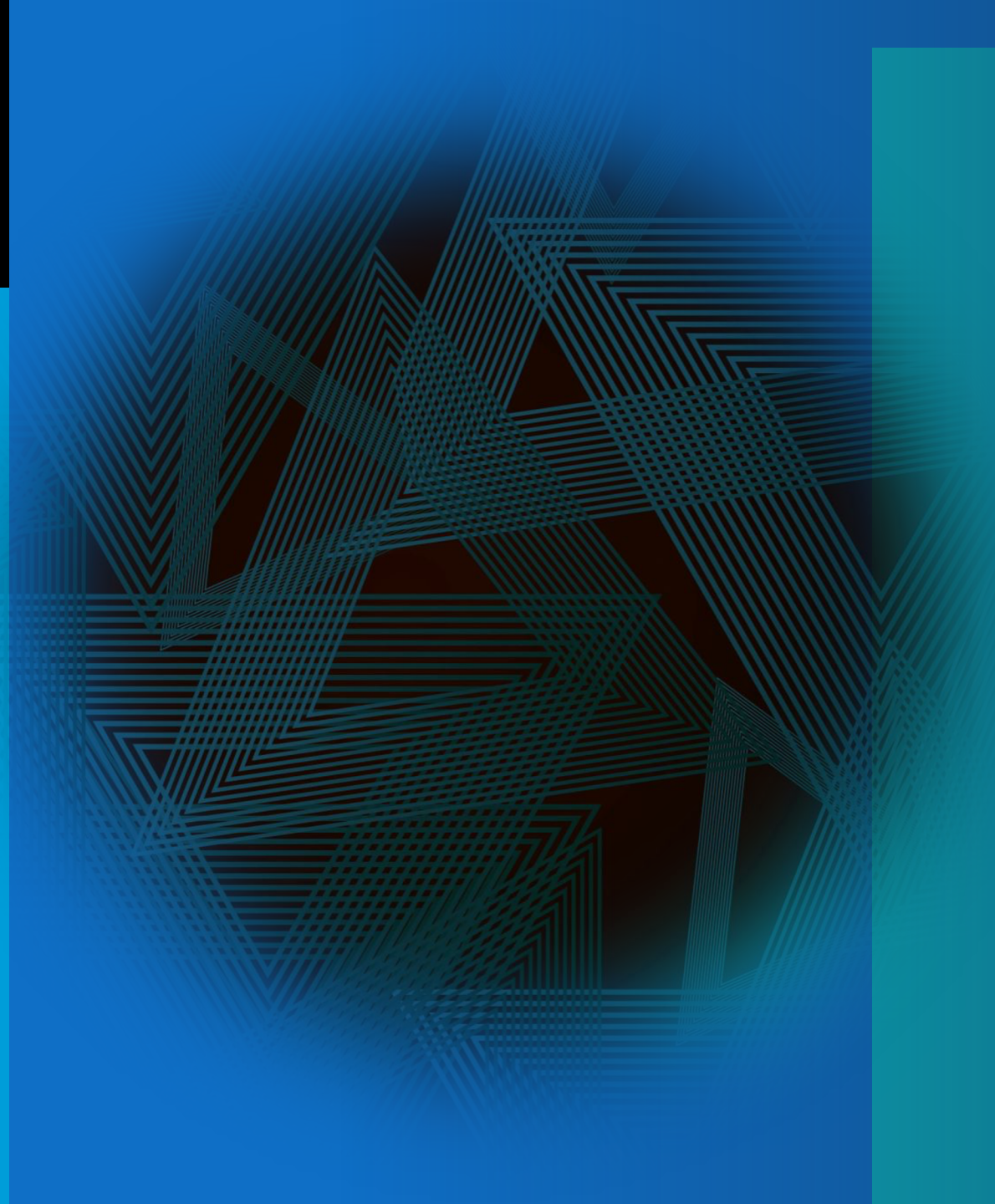


Présentation d'un article scientifique

Text Detoxification using Large Pre-trained Neural Models

ADVANCED MACHINE LEARNING FOR NLP AND TEXT PROCESSING

YANIS FADILI & LIES HAOUAS
DIA 2

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Sommaire

Contexte de l'article

Skoltech
Skolkovo Institute of Science and Technology

MTS

Text Detoxification using Large Pre-trained Neural Models

David Dale[‡], Anton Voronov[‡,†], Daryna Dementieva[‡], Varvara Logacheva[‡], Olga Kozlova[†], Nikita Semenov[†], and Alexander Panchenko[‡]
[‡]Skolkovo Institute of Science and Technology, Moscow, Russia
[†]Mobile TeleSystems (MTS), Moscow, Russia
{d.dale,anton.voronov,daryna.dementieva,v.logacheva,a.panchenko}@skoltech.ru
{oskozlo9,nikita.semenov}@mts.ru

Abstract

We present two novel unsupervised methods for eliminating toxicity in text. Our first method combines two recent ideas: (1) guidance of the generation process with small style-conditional language models and (2) use of paraphrasing models to perform style transfer. We use a well-performing paraphraser guided by style-trained language models to keep the text content and remove toxicity. Our second method uses BERT to replace toxic words with their non-offensive synonyms. We make the method more flexible by enabling BERT to replace mask tokens with a variable number of words. Finally, we present the first large-scale comparative study of style transfer models on the task of toxicity removal. We compare our models with a number of methods for style transfer. The models are evaluated in a reference-free way using a combination of unsupervised style transfer metrics. Both methods we suggest yield new SOTA results.

1 Introduction

Identification of toxicity in user texts is an active area of research (Zampieri et al., 2020; D'Sa et al., 2020; Han and Tsvetkov, 2020). The task of automatic rewriting of offensive content attracted less attention, yet it may find various useful applications such as making online world a better place by suggesting to a user posting a more neutral version of an emotional comment. The existing works on text detoxification (dos Santos et al., 2018; Tran et al., 2020; Laugier et al., 2021) cast this task as style transfer. The style transfer task is generally understood as rewriting of text with the same content and with altering of one or several attributes which constitute the "style", such as authorship (Voigt et al., 2018), sentiment (Shen et al., 2017), or degree of politeness (Madaan et al., 2020). Despite the goal of preserving the content, in many cases changing the style attributes changes the meaning of a sentence significantly.[1] So in fact the goal of many style transfer models is to transform a sentence into a somewhat similar sentence of a different style on the same topic.[2] We suggest that detoxification needs better preservation of the original meaning than many other style transfer tasks, such as sentiment transfer, so it should be performed differently.

We present two models for text detoxification, which have extra control for content preservation. The first model, **ParaGeDi**, is capable of fully regenerating the input. It is based on two ideas: external control of an output of a generation model by a class-conditioned LM (Krause et al., 2020) and formulation of style transfer task as paraphrasing (Krishna et al., 2020). Being based on a paraphraser model, **ParaGeDi** explicitly aims at preserving the meaning of the original sentence. The second approach, **CondBERT**, inspired by Wu et al. (2019a), follows the pointwise editing setup. It uses BERT to replace toxic spans found in the sentence with their non-toxic alternatives. The semantic similarity is maintained by showing the original text to BERT and reranking its hypotheses based on the similarity between the original words and their substitutes. Interestingly, BERT does not need any class-conditional pre-training to successfully change the text style from toxic to normal.

In addition, we perform a large-scale evaluation of style transfer models on detoxification task, comparing our new models with baselines and state-of-the-art approaches. We release our code and data.[3]

Our contributions are as follows:

- We propose two novel detoxification methods based on pre-trained neural language models: **ParaGeDi** (paraphrasing GeDi) and **CondBERT** (conditional BERT).

[1]For example, Lample et al. (2019) provide the following sentence as an example of transfer from male to female writing: *Gotta say that beard makes you look like a Viking → Gotta say that hair makes you look like a Mermaid.*
[2]A formal task definition is presented in Appendix A.
[3]https://github.com/skoltech-nlp/detox

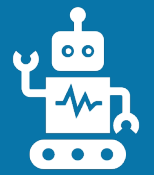
La "text-detoxification"

« **Méthode** consistant à **réécrire un texte**, tout en **préservant** sa signification (le **fond**) et en **retirant** son style dit toxique (la **forme**). »

« Je suis à la bourre pour aller au boulot ! 😠 »

↓

« Je risque d'arriver en retard au travail. 🏃‍♀️ »



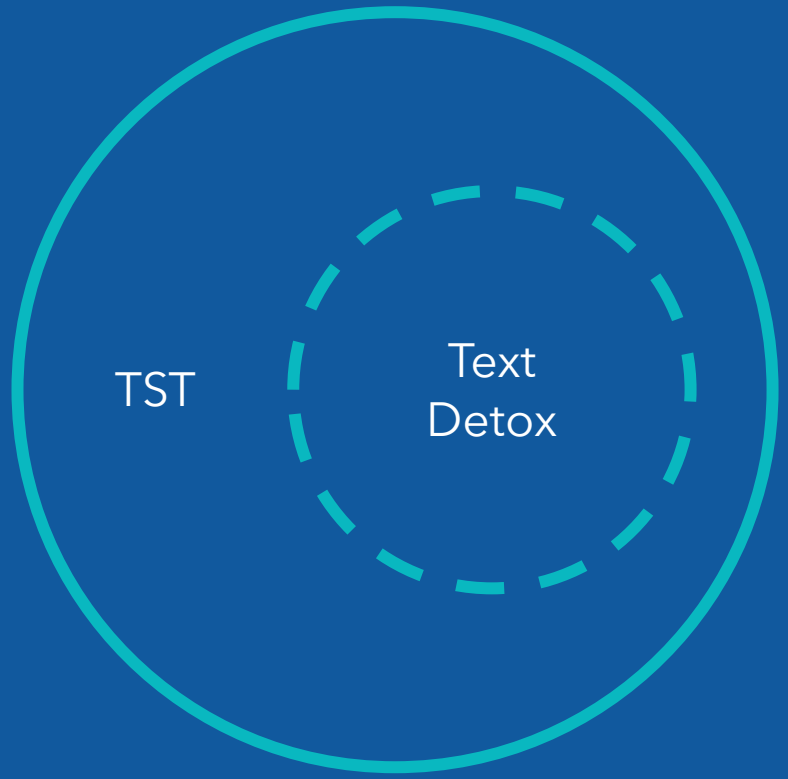
Chatbots

Nombreux domaines d'application

Assistants vocaux

Réseaux sociaux

Le TST (Text Style Transfer)

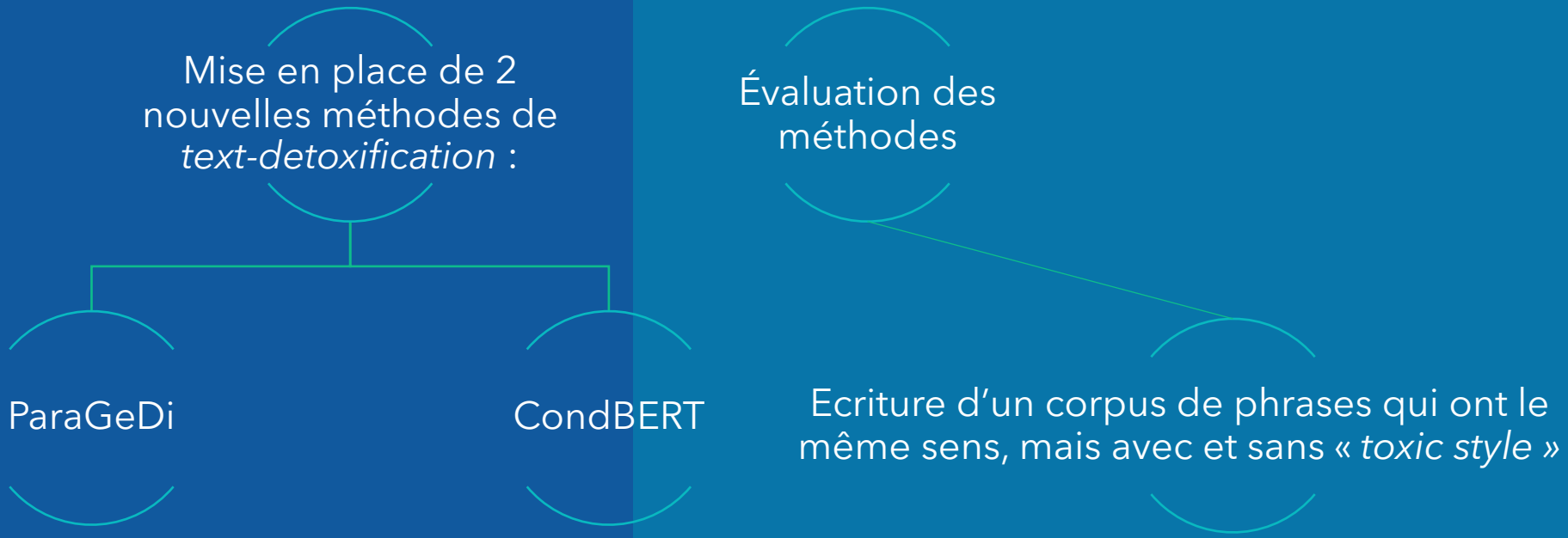


TST
- Contenu (fond)
- Style (forme)

→ Concepts assez ambigus → Peuvent changer le sens d'une phrase

- *Unsupervised* style transfer:
  - There are no parallel corpora for direct translation from toxic to non-toxic
  - There are large non-parallel corpora labelled by style
  - There are large paraphrase corpora without style labels

Contribution des auteurs



État de l’art & travaux liés

Dans le transfert de style, on relève 3 catégories de modèles :

Disentangled latent representation

- Problème : la plupart des méthodes ne sont pas efficace en termes de préservation du sens des phrases

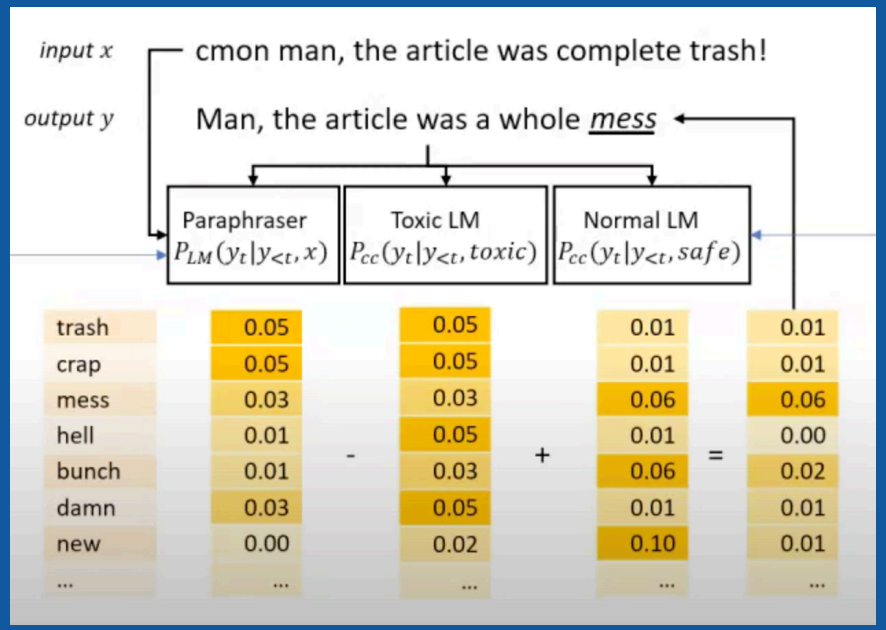
Pointwise edition

- Delete-Retreive-Generate : remplace les mots d'autres mots en se basant sur l'utilisation des ces mots dans d'autres phrases
- Mask-and-Infill : remplace des mots en utilisant la méthode MLM (Masking Language Models)

Back-translation and cycle consistency

- STRAP : retirer le style des phrases en les paraphrasant puis en entraînant le modèle sur ces paraphrases pour refaire la phrase initiale

1ère alternative : ParaGeDi



Consiste à générer un texte à partir de zéro, mais guidé par un modèle de langue entrainé sur des aspects spécifiques du texte tels que le style, le thème, etc. Les auteurs décrivent formellement ce modèle grâce des formules mathématiques définissant le modèle de génération.

Afin de permettre à GeDi de conserver le sens du texte d'entrée, ils ont remplacé le modèle langage de base par un modèle capable de paraphraser.

Le processus consiste à détecter les mots toxiques dans l'entrée, puis la paraphraser en choisissant parmi les mots toxiques ceux qui respectent une formule mathématique calculant un compromis entre les paraphrases les plus probables avec les mots les moins toxiques.

2ème alternative : CondBERT

Consiste à générer un texte à partir de zéro, mais guidé par un modèle de langue entrainé sur des aspects spécifiques du texte tels que le style, le thème, etc. Les auteurs décrivent formellement ce modèle grâce des formules mathématiques définissant le modèle de génération.

Afin de permettre à GeDi de conserver le sens du texte d'entrée, ils ont remplacé le modèle langage de base par un modèle capable de paraphraser.