TOOPLOOX

Dominika Basaj

WILDNLP

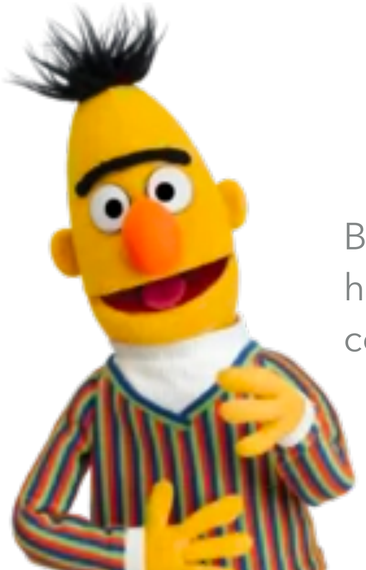
An open-source framework for making sure your NLP models work in the wild.

NLP is cracked … almost!

NLP is booming since 2018

Models surpass humans

But humans can break them!



BERT & his relatives have beaten humans a couple of times…

ELMo - started anti-human revolution

Who hasn't been there?

NLP models are easy to break.

Robots can take on any form but some are made to resemble humans in appearance. This is said to help in the **acceptance** of a robot in certain replicative behaviors usually performed by people. Such robots **attempt to replicate walking, lifting, speech, cognition, and basically anything a human can do.**

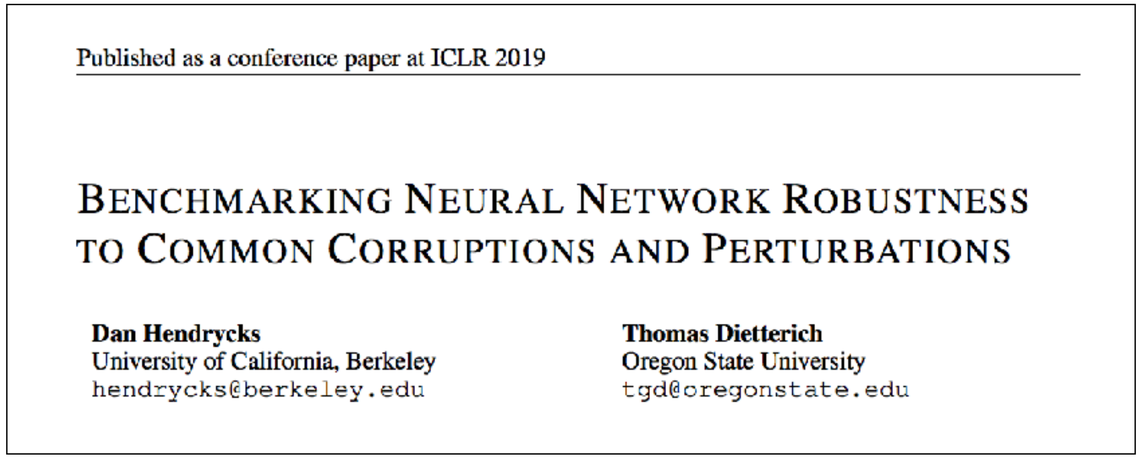
Waht do robots that resemble humans attempt to **so**?

acceptance!

Ugh? **What** do robots that resemble humans attempt to **do**?

Oh, ok, now I got you! Attempt to replicate walking, lifting, speech, cognition, and basically anything a human can do

NLP is a little bit behind...



Published as a conference paper at ICLR 2019

BENCHMARKING NEURAL NETWORK ROBUSTNESS TO COMMON CORRUPTIONS AND PERTURBATIONS

Dan Hendrycks
University of California, Berkeley
hendrycks@berkeley.edu

Thomas Dietterich
Oregon State University
tgd@oregonstate.edu

NLP has no unified robustness benchmark

Model comparison is not possible!

WildNLP: an overview

WildNLP

Designed attacks

- Alleviates the problem
- Facilitates error-prone training

Generation of natural errors

- Based on corpus with natural errors

Adversarial robustness training

- Showing erroneous examples is not enough

Designed perturbations as a first step.

Aspect	Example sentence
Original	Warsaw was believed to be one of the most beautiful cities in the world.
`Article`	Warsaw was believed to be one of **a** most beautiful cities in world.
`Swap`	Warsaw **aws** believed to be one **fo teh** most beautiful cities in the world.
`Qwerty`	Wa**d**saw was b**d**lieved to be one of the most beautiful citie**e** in the world.
`Remove_char`	Warsaw was believed to be one **o th** most **eautiful** cities in the world.
`Remove_space`	Warsaw was believed **tobe** one of the most beautiful cities in the world.
Original	You cannot accidentally commit vandalism. It used to be a rare occurrence.
`Misspelling`	You can not **accidentaly** commit vandalism. It used to be a rare **occurrance**.
Original	Bus Stops for Route 6, 6.1
`Digits2words`	Bus Stops for Route **six**, **six point one**
Original	Choosing between affect and effect can be scary.
`Homophones`	Choosing between **effect** and effect can **bee** scary.
Original	Laughably foolish or false: an absurd explanation.
`Negatives`	**Laughab*y fo*lish** or **fal*e**: an **a*surd** explanation.
Original	Sometimes it is good to be first, and sometimes it is good to be last.
`Positives`	Sometimes it is **go*d** to be first, and sometimes it is **goo*** to be last.
`Marks`	Sometimes, it is good to be first and sometimes, it, is good to be last.

Soon!

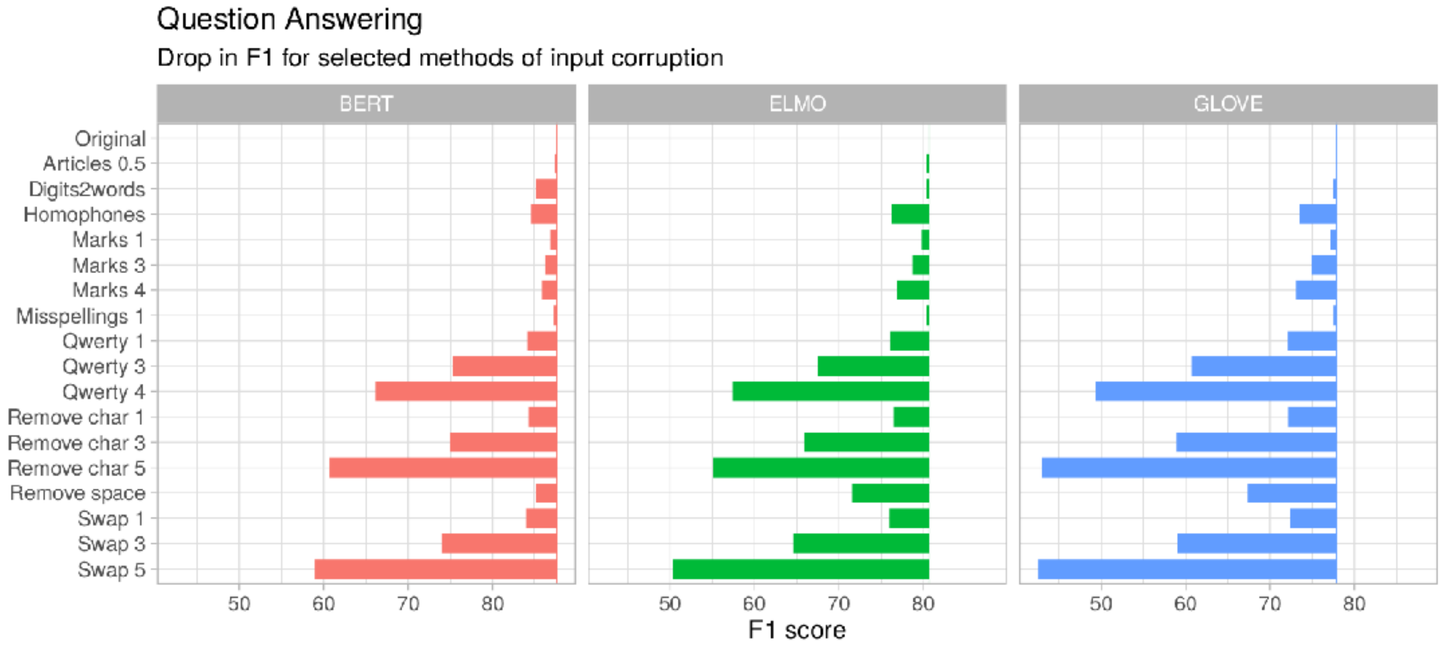
Generation of natural errors

Soon!

Adversarial robustness training

Fragile? Should I be worried?

Yes, you should!



- Extreme overfitting to training set
- Datasets are too specific
- Great performance does not equal generalization

WildNLP

11k downloads

https://github.com/MI2DataLab/WildNLP

pip install wild-nlp

Thank you!
Let’s stay in touch!

dominika.basaj@tooploox.com

TOOPLOOX AI