

Breaking news in NLP: bigger is better... but is it really?

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAcCT*.

Language models

A system train on **string prediction tasks**.

Language models

A system train on **string prediction tasks**.

Do not miss our session tomorrow!

Language models

A system train on **string prediction tasks**.

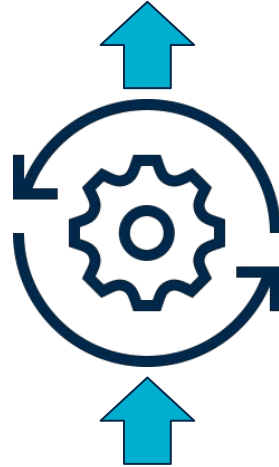
Do not miss our <?> tomorrow!



Language models

A system train on **string prediction tasks**.

session 3% - sale 2.5% - webinar 1% [...]



Do not miss our <?> tomorrow!

Introduction

Language models

The historic way

word2vec (2013)

GloVe (2014)

context2vec (2019)

ELMo (2018)

Language models

Some new challengers

Model	# of parameters	Dataset size	Number of books
BERT	3.4E+08	16GB	16,000
MegatronLM	8.30E+09	161GB	161,000
GPT-3	1.75E+11	570GB	570,000
Switch-C	1.57E+12	745 GB	745,000




A new paradigm: **pre-training + finetuning.**







Language

To new heights

Rank Name	Model	URL Score
1 DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	 90.8

	RankName	Model	URL Score E
+	1 DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4	 90.3
+	2 Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)	90.2
	3 SuperGLUE Human Baselines	SuperGLUE Human Baselines	 89.8
+	4 T5 Team - Google	T5	 89.3

12	Junjie Yang	HIRE-RoBERTa	 88.3
13	Facebook AI	RoBERTa	 88.1
+	14 Microsoft D365 AI & MSR AI	MT-DNN-ensemble	 87.6
	15 GLUE Human Baselines	GLUE Human Baselines	 87.1



WHY ARE SOME PEOPLE COMPLAINING?

Training cost

Environmental cost

These models require a **huge computing power**.

Strubell et al. (2019) estimated that training BERT was roughly equivalent to a **trans-American flight**.

And that is only the training part! **You need to add energy usage at each inference!**



Financial cost

Energy is not free, and huge usage of energy means **huge expenses**.

Strubell et al. (2019) estimated it will cost between 4,000 and 12,000 \$ to train one BERT on a cloud computing platform.

For a company **the inference cost can easily surpass the training cost**.



More ethical issues

Not an equal access: need computing power to use them that not everybody have.

Negative environmental impact

Risks / benefits are not well distributed: the first communities impacted by climate change are not the one benefiting these models.

Uncurated dataset

Diversity in the data

Data comes from **web crawling**:

- Internet contributors are **not evenly distributed**
- Marginalized populations **avoid mainstream sites**
- **Hegemonic viewpoint are overrepresented** even compared to the general population
- **Documentation debt**

⇒ **We can't consider this data as 'representative' of all humanity**



Models are even more biased

It has been proven that language models **encode even more bias than the data**:

- BERT tends to associate disabled persons with more negative sentiment words
- GPT-3 can generate toxic texts even when prompted with non-toxic text (in GPT-2 documents some documents of the dataset were coming from banned sub-reddit)

As we all come with our own biases, we **need to work with marginalized population** to audit for bias.

⇒ **Impossible to do it for hundred of GB of text!**



Models are even more biased

**“Feeding AI systems on the world’s beauty, ugliness, and cruelty,
but expecting it to reflect only the beauty is a fantasy.”**

Abeba Birhane and Vinay Uday Prabhu. 2021. Large Image Datasets: A Pyrrhic Win for Computer Vision?. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

1537–1547



The easy path

Easy path

Currently, people are saying language models will lead us to **natural language understanding**.

It overshadows other research trying to make systems work on small curated datasets.

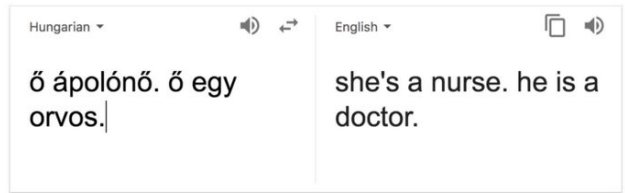
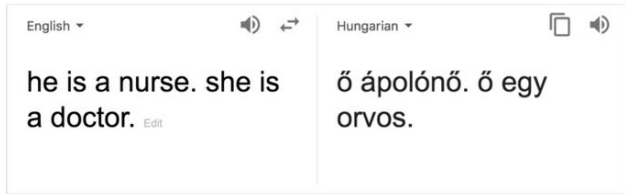
Language is pairing forms and meaning but language models only have access to form.



SOME REAL LIFE EXAMPLES

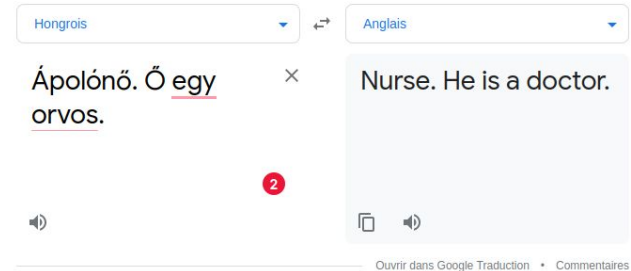
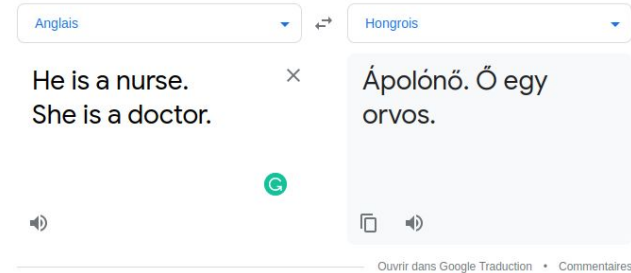
Gender bias in translation

My doctor is a she and my nurse a he.



05/12/2017:

<https://medium.com/@laurahelendouglas/ai-is-not-just-learning-our-biases-it-is-a-mplifying-them-4d0dee75931d>



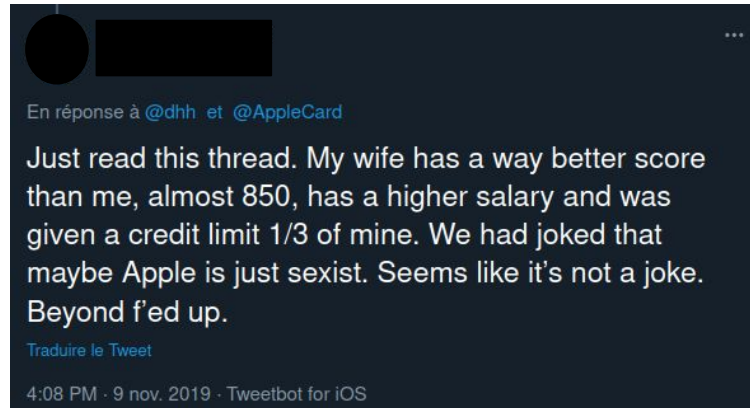
Test made 24/02/21

⇒ Easy to reproduce with language where name doesn't have grammatical gender like

Turkish or Hungarian.

Gender bias with the Apple Card

“But we don’t have the data about the gender!”



Gender bias with the Apple Card

“But we don’t have the data about the gender!”

Goldman landed on what sounded like an ironclad defense: The algorithm, it said, has been vetted for potential bias by a third party; moreover, **it doesn’t even use gender as an input. How could the bank discriminate if no one ever tells it which customers are women and which are men?**

<https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/>



Lost in translation

Thanks Facebook

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

<https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest>

24/10/17

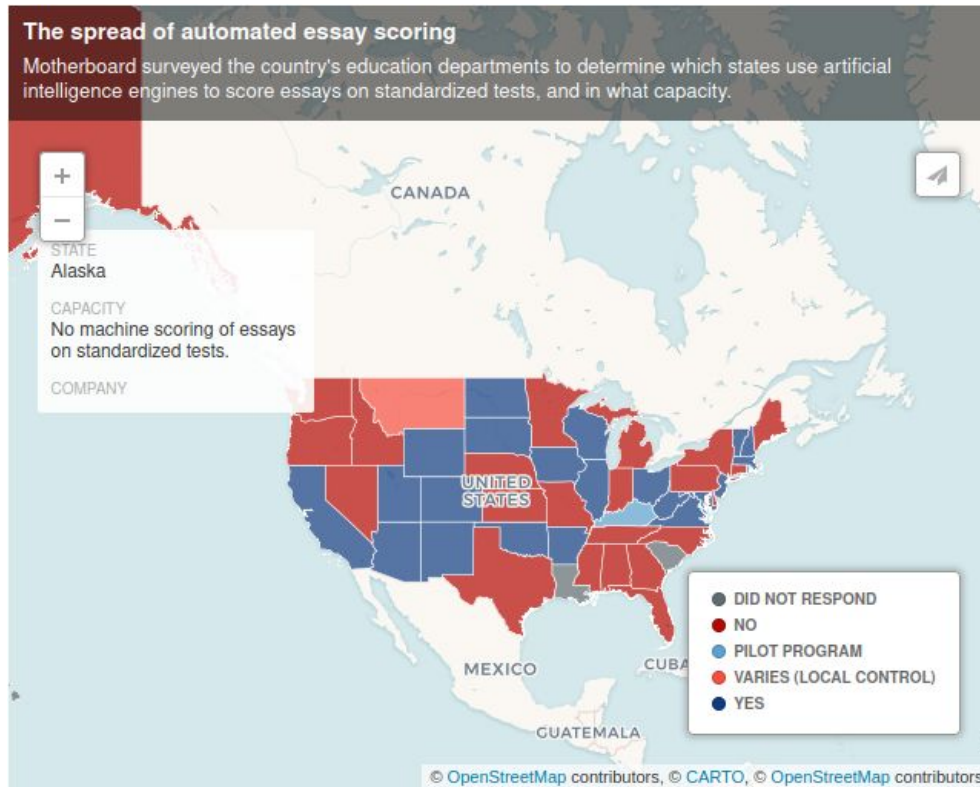
Base message: “good morning” (in Arabic)

To: “hurt them” (in English)

To: “attack them” (in Hebrew)

Automated essay grading

A dream of many



<https://www.vice.com/en/article/pa7di9/flawed-algorithms-are-grading-millions-of-students-essays>

Automated essay grading

Review

Analyze an Issue Topic:

In most professions and academic fields, imagination is more important than knowledge.

Write a response in which you discuss the extent to which you agree or disagree with the claim. In developing and supporting your position, be sure to address the most compelling reasons and/or examples that could be used to challenge your position

Your Answer:

Careers with corroboration has not, and in all likelihood never will be compassionate, gratuitous, and disciplinary. Mankind will always proclaim noesis; many for a trope but a few on executioner. a quantity of vocation lies in the study of reality as well as the area of semantics. Why is imaginativeness so pulverous to happenstance? The reply to this query is that knowledge is vehemently and boisterously contemporary.

[The BABEL Generator and E-Rater: 21st Century Writing Constructs and Automated Essay Scoring \(AES\). Journal of Writing Assessment 13:1 \(2020\)](#)

Google believes in ethics ...

2018: creation of the Google's Ethical AI team lead by Margaret Mitchell

2020: Google's Ethical AI team co-lead by Timnit Gebru and Margaret Mitchell

Jeff Dean (@JeffDean)
I understand the concern over Timnit's resignation from Google. She's done a great deal to move the field forward with her research. I wanted to share the email I sent to Google Research and some thoughts on our research process.

[Traduire le Tweet](#)

[About Google's approach to research publication](#)
About Google's approach to research publication I understand the concern over Timnit Gebru's resignation from Google. She...
[docs.google.com](#)

9:12 PM · 4 déc. 2020 · Twitter Web App

Timnit Gebru
I was fired by @JeffDean for my email to Brain women and Allies. My corp account has been cutoff. So I've been immediately fired :-)

[Traduire le Tweet](#)

5:24 AM · 3 déc. 2020 · Twitter Web App

MMitchell
I'm fired.

11:02 PM · 19 févr. 2021 · Twitter Web App

IMPROVEMENTS

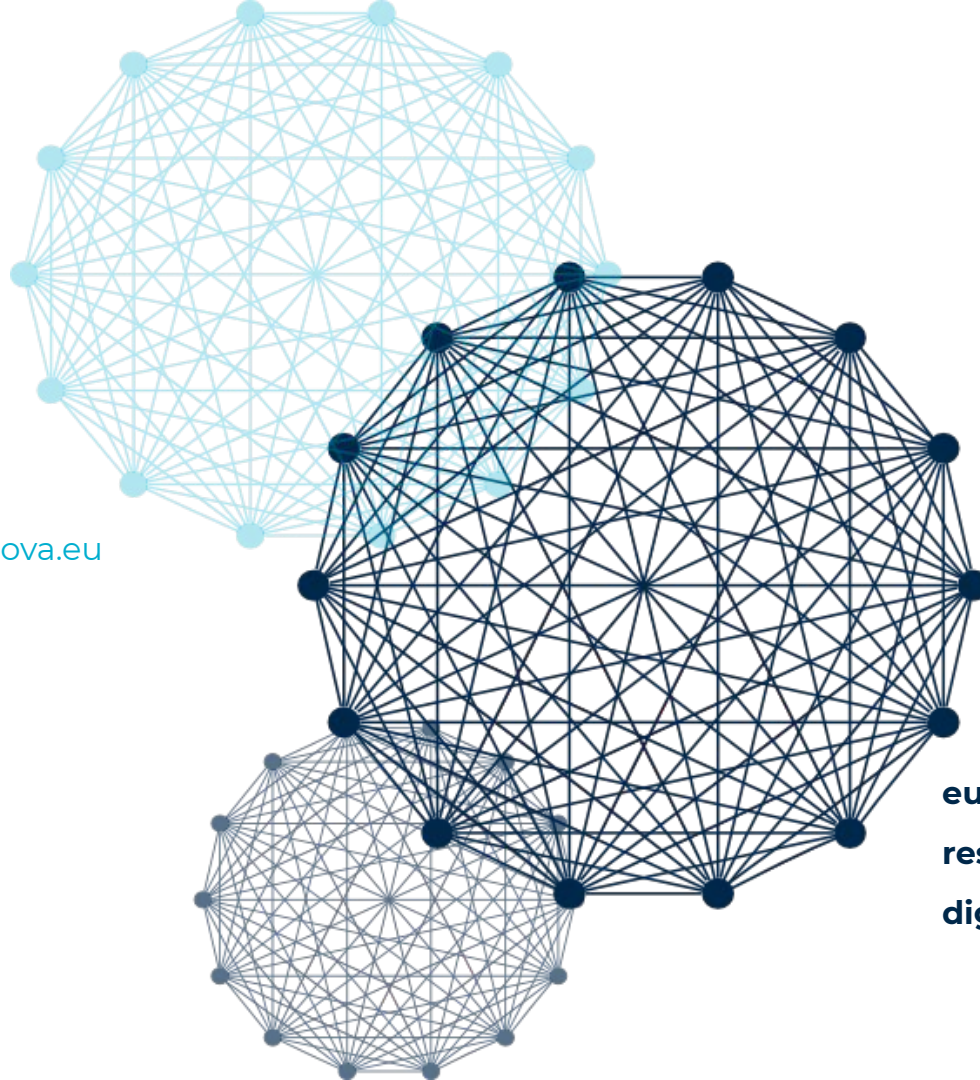
What can we do?

“We should consider our research time and effort a valuable resource, to be spent to the extent possible on research projects that build towards a technological ecosystem whose benefits are at least evenly distributed or better accrue to those historically most marginalized.”

Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAccT*.

What can we do?

- From making bigger models to understanding of the models are achieving the tasks.
- Frameworks to describe the uses of your models.
- Identify who will be impacted directly and indirectly by your model and work with them.
- Thinking more about the trade-off between energy consumption and metric improvement.



Contact:

leo.bouscarrat@euranova.eu

euranova.eu

research.euranova.eu

digazu.com

Resources

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). [On the dangers of stochastic parrots: Can language models be too big.](#) In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA.

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). [Energy and Policy Considerations for Deep Learning in NLP.](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650).

[A critic of Bender and Gebru et al. 2019 by Yoav Goldberg.](#)

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). [Model cards for model reporting.](#) In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).