On the Multilingual Capabilities of Very Large-Scale English Language Models

Jordi Armengol-Estapé[1], **Ona de Gibert**[1] **& Maite Melero**

Equal contribution[1]

Barcelona Supercomputing Center, Barcelona, Spain[1]

jordi.armengol@bsc.es

@jordiae



Abstract

Generative Pre-trained Transformers (GPTs) have recently been scaled to unprecedented sizes in the history of machine learning. These language models have been shown to exhibit outstanding zero, one, and few-shot learning capabilities in a number of different tasks. Nevertheless, aside from anecdotal experiences, little is known regarding their multilingual capabilities, given the fact that the pre-training corpus is almost entirely composed of English text. In this work, we investigate its potential and limits in three tasks: extractive Question-Answering, text summarization and natural language generation for five different languages, as well as the effect of scale in terms of model size. Our results show that GPT-3 can be used, not only as a powerful generative pre-trained model for English, but for other languages as well, even for some with very few data in the training corpora, with room for improvement if optimization of the tokenization is addressed.

Introduction

The arrival of GPT-3 [1]:

- Biggest (non-sparse, trained until convergence) language model ever at the time of publication.
- Trained large dataset - mainly for English, 93% by word count, plus anecdotal presence other languages.

GPT-3 exhibits outstanding NLU/NLG capabilities in English - but could it also work for other languages as well? If it does, it would be a phenomenal proof of transfer learning, because GPT-3 is basically a monolingual model.

In these experiments, we investigate the multilingual skills of different size variants of the GPT-3 model.

Related Work

- Increase in performance with model size, data and compute ([2])
- Evaluation of GPT-3 in several tasks with scaling and zero and few-shot settings for English [1]
- Experiments with GPT-3 generative capabilities in English [3] [4] [5]
- Ethical concerns of GPT-3 [6] [7]
- How to optimally prompt the model [8]

No other work has systematically studied its potential for solving tasks in languages other than English, aside from machine translation.

Methodology

We use OpenAI's API with:

- 4 model sizes 1. Ada (350M) 2. Babbage (1.3 B) 3. Curie (6.7B) 4. Davinci (175B)
- 3 tasks 1. Question Answering 2. Text Summarization 3. Text Generation
- 6 languages 1. Catalan 2. German 3. Spanish 4. Russian 5. Turkish 6. + 6. English, as reference
- 2 evaluations 1. Automatic metrics (F1, ROUGE...) 2. Human evaluation

Prompting (zero-shot):

This is a question-answering system in English.
Context: The Panthers defense gave up just 308 points [...]
Question: How many points did the Panthers defense surrender?
Answer: ***308***

Results

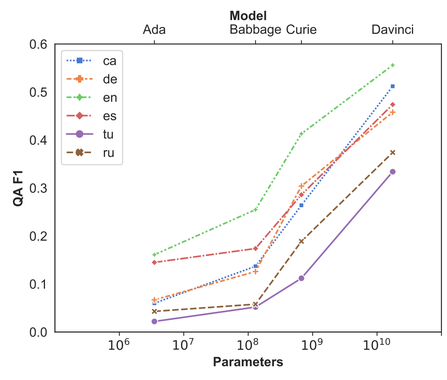
Tokenization

Average tokens per word in the studied datasets:

	CA	DE	EN	ES	RU	TU
Summarization	2.13	2.43	1.23	1.98	-	3.61
Question Answering	2.12	2.68	1.290	2.06	7.96	3.66

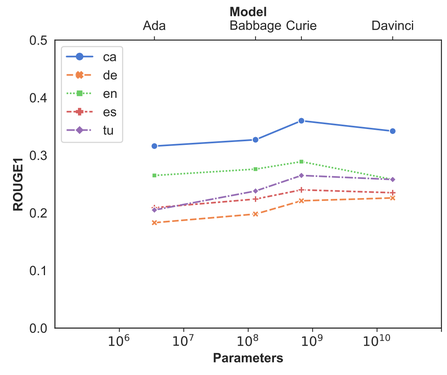
Zero-shot Question answering

XQuAD [9] (+ XQuAD ca[10]) results:



Zero-shot Summarization

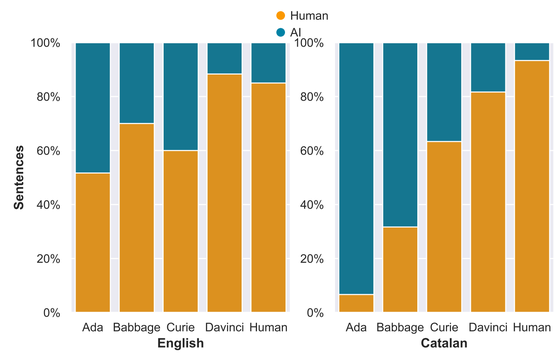
CNN Daily Mail [11] + MLSum [12] + CaSum [13]:



English	1st	2nd	3rd	4th	5th
Ada	8	15	21	23	33
Babbage	13	21	23	27	16
Curie	15	27	20	17	21
Davinci	16	19	25	17	23
Human	49	19	13	12	7

Catalan	1st	2nd	3rd	4th	5th
Ada	9	12	23	16	40
Babbage	7	27	15	31	21
Curie	12	19	33	25	11
Davinci	21	17	19	20	23
Human	51	25	13	5	5

Unconditional Generation



Discussion

- Results:
  - Question answering: Steep scaling curve.
  - Summarization: The challenging evaluation affects the study, but GPT-3 is remarkably good at e.g. Catalan sumarization.
  - Unconditional generation: Scaling is more noticeable in Catalan than in English because small English LMs are already decent at generating sentences.
- Tokenization: Word per token is indeed useful for predicting GPT-3 performance for a given language; Russian summarization is not even possible.
- Scaling: Transfer learning between English and the other languages in zero-shot settings scales with model size in a very steep curve
- Usability in practice affected by the tokenization-dependent pricing.
- Limitations of our study: Evaluating generative tasks is hard. We do not have control over the used models, so we cannot study different tokenizers, model sizes or data. We cannot fit scaling laws due to the lack of data points.

Conclusions

- The study of how scaling affects multilingual performance could allow to forecast multilingual performance of future English language models.
- Multilingual capabilities of large English language models: In spite of the tiny multilingual data in the train corpus and the English-centric tokenizer, GPT-3 exhibits remarkable zero-shot multilingual capabilities. Results are surprisingly close to the reference results for English

  This confirms the extraordinary capacity of massive LMs to generalise not only across tasks but also languages, acting as a universal interlingua.
- Future work: Extend the study of the scaling laws in LMs for cross-lingual transfer.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020.

[3] Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*, 2021.

[4] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[5] Katherine Elkins and Jon Chun. Can gpt-3 pass a writer's turing test. *Journal of Cultural Analytics*, 2371:4549, 2020.

[6] Robert Dale. Gpt-3: What's it good for? *Natural Language Engineering*, 27(1):113–118, 2021.

[7] Kris McGuffie and Alex Newhouse. The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*, 2020.

[8] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *CoRR*, abs/2101.06804, 2021.

[9] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2019.

[10] Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online, August 2021. Association for Computational Linguistics.

[11] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015.

[12] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. Mlsum: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, 2020.

[13] Ona de Gibert, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estapé, and Maite Melero. Sequence-to-sequence resources for catalan, 2022.

Acknowledgements

This work was funded by the MT4All CEF project.