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

Jordi Armengol-Estapé¹, Ona de Gibert¹ & Maite Melero²
Equal contribution¹
Barcelona Supercomputing Center, Barcelona, Spain¹
jordi.armengol@bsc.es @jordia

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 about 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
• 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:

<table>
<thead>
<tr>
<th>Language</th>
<th>CA</th>
<th>DE</th>
<th>EN</th>
<th>ES</th>
<th>RU</th>
<th>TU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summarization</td>
<td>2.13</td>
<td>2.65</td>
<td>2.73</td>
<td>1.90</td>
<td>3.61</td>
<td></td>
</tr>
<tr>
<td>Question Answering</td>
<td>2.12</td>
<td>2.68</td>
<td>2.00</td>
<td>7.96</td>
<td>3.66</td>
<td></td>
</tr>
</tbody>
</table>

Zero-shot Question answering

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

<table>
<thead>
<tr>
<th>Language</th>
<th>CA</th>
<th>DE</th>
<th>EN</th>
<th>ES</th>
<th>TU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Zero-shot Summarization


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 summarization.
 – 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


Acknowledgements

This work was funded by the MT4All CEF project.