Natural Language Generation with Architecture of Transformers: A Case Study of Business Names Generation

Mantas Lukauskas[1,2] Tomas Rasymas[1], Domas Vaitmonas[1], Matas Minelga[1]

1) Zyro Inc., Kaunas, Lithuania (www.zyro.com)

2) Faculty of Mathematics and Natural Sciences, Kaunas University of Technology

mantas.lukauskas@ktu.lt / mantas.lukauskas@zyro.com

Abstract

The continuous improvement of artificial intelligence/machine learning leads to an increasing search for the broader application of these technological solutions to structured and unstructured data. One of the applications for unstructured data is natural language processing (NLP). Natural language processing is finding more and more different ways to adjust to real practical problems. These tasks can range from finding meaningful information in unstructured data, analysing sentiments, and translating the text into another language to fully automated human-level text creation. This study aims to apply natural language modelling models and the architecture of transformers to generate high-quality business names.

Data and Methods

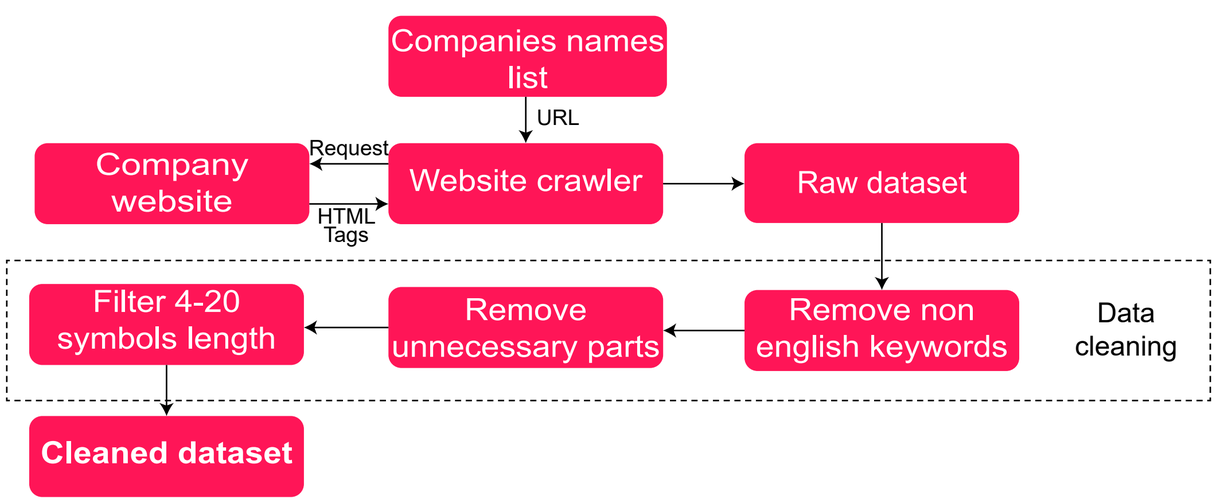

Figure 1. Data gathering process

The dataset for this study consists of 350,928 business names. This data was collected using the websites of start-ups from all over the world. The experiments in this study were performed using a Google Cloud Platform virtual machine with parameters:12 vCPUs, 78 GB RAM, 1 x NVIDIA Tesla T4 GPU (16 GB VRAM). For the biggest models, virtual machine parameters have been increased to 16vCPUs, 150GB of RAM, and 2x NVIDIA Tesla T4. To speed up the calculations, the Python Facebook DeepSpeed library was used. DeepSpeed uses ZeRO (Zero Redundancy Optimizer) optimization strategies. These strategies eliminate excess memory in all parallel data processes by dividing the three model states (optimizer states, slopes, and parameters) in parallel data processes rather than repeating them.

Results

A critical element in using the ZeRO3 optimizer is the high RAM usage. The chart below provides information on the use of RAM in training different models. It can be seen that the highest RAM usage is observed in the most significant model GPT-J-6B. This usage is as high as 101 GB. It is also noted that GPT2-XL and GPTNeo-1.3B have a quite similar RAM usage. The interesting fact is that the GPT model uses more RAM compared to GPT2 and DistilGPT2.

Model	RAM/VRAM	Perplexity	Avg. Score (SD)
ADA*	—	—	41.81 (10.05)
Babbage*	—	—	37.86 (9.35)
Curie*	—	—	35.92 (8.07)
GPT	10.6/15	2.46	38.88 (10.76)
DistilGPT2	9.5/15	9.49	39.67 (10.61)
GPT2	10.5/14.5	10.26	43.25 (11.42)
GPT2-M	14.7/14.8	8.18	42.23 (12.45)
GPT2-L	22.7/14.9	10.86	45.66 (11.08)
GPT2-XL	34.8/14.9	17.62	42.71 (11.41)
Neo-125M	10.6/15	9.12	44.66 (10.75)
Neo-1.3B	31.7/14.8	36.37	46.60 (11.36)
Neo-2.7B	49/12.7	41.62	44.93 (9.81)
GPT-J-6B	101/15	37.08	—

** Model tested on OpenAI environment due to that do not have RAM/VRAM and Perplexity*

Figure 2. Comparison of different NLP model used in research

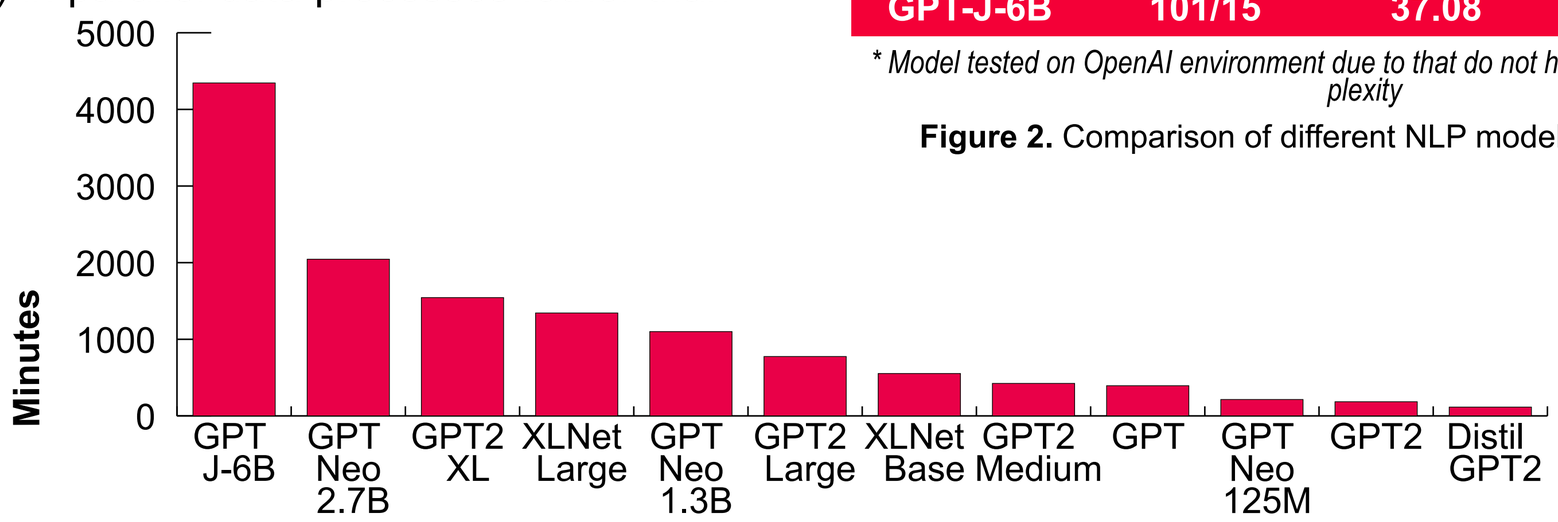

Figure 3. Model fine-tune time with Testa T4 GPU (minutes)

Conclusion

The obtained results showed that in the case of business name generation, the larger models do not have statistically significantly better results compared to the smaller models. Aapplication of larger models in practice is not beneficial because the generation of larger model names takes a statistically significant longer time than the generation of names with smaller models. It is also noticeable that the new generation of transformers features much better generation of business names. The results of the study show that people's assessment and assessment by perplexity are different. In human evaluation, it is observed that the best result is obtained using the GPT-Neo-1.3B model. The evaluation of this model is statistically significantly higher compared to other models ($p < 0.05$). Interestingly, the GPT-Neo-2.7B model has poorer results. Its evaluation does not differ statistically significantly from the GPT-Neo-125M model ($p > 0.05$), which is even 20 times smaller. A critical element in using the ZeRO3 optimizer is the high RAM usage. The highest RAM usage is observed in the most significant model GPT-J-6B. This usage is as high as 101 GB. It is also noted that GPT2-XL and GPT-Neo-1.3B have a pretty similar RAM usage. The interesting fact is that the GPT model uses more RAM compared to GPT2 and DistilGPT2.

Test It in Production

https://zyro.com/tools/business-name-generator

Input

Pancakes, Bakery

Output

Cupcakes Mrs, The Cookie Dough, Bakery Roots, Mere Bakery, Cupcakes eatery