Transformers architecture application in high-quality business names generation

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

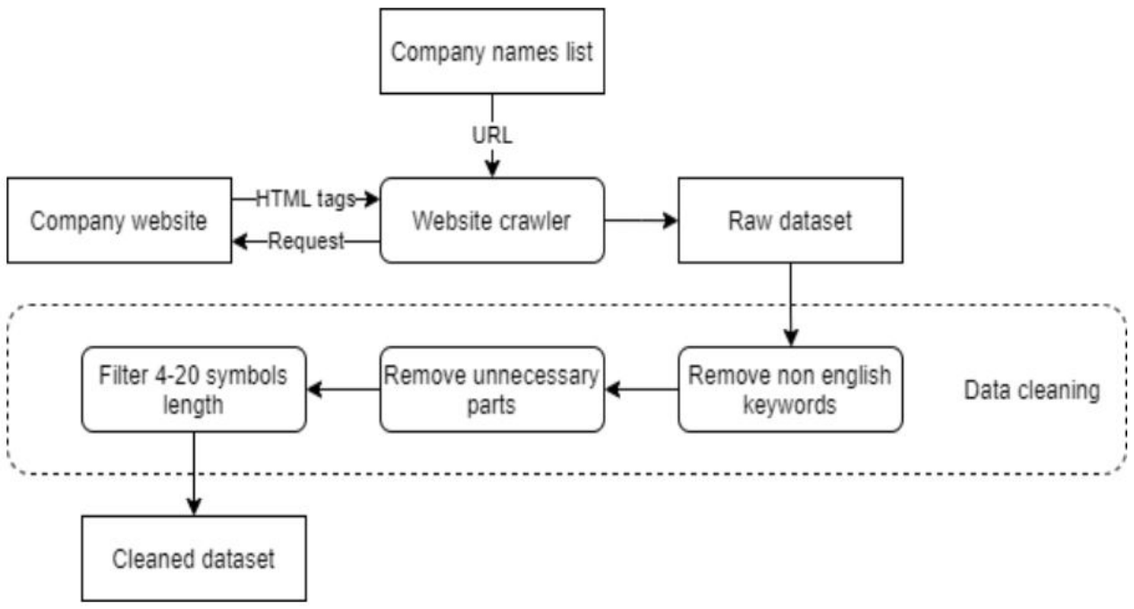
1) Zyro Inc., Kaunas, Lithuania (www.zyro.com)
2) Faculty of Mathematics and Natural Sciences, Kaunas University of Technology

mantas.lukauskas@ktu.lt

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 the computer analysis and processing of natural language (which can be delivered and written) using various technologies. NLP aims at linguistically adapted various tasks or computer programs in human languages. 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. The dataset for this study consists of 350,928 observations/business names (299,964 training and 50,964 observations in the test sample). This data was collected using the websites of start-ups from all over the world. For different models comparison, the data set was divided into two parts. The training data set represented 80%, and the test data set 20%. The experiments in this study were performed using a Google Cloud Platform virtual machine with parameters:12 vCPUs, 78 GB random access memory (RAM), 1 x NVIDIA Tesla T4 GPU (16 GB VRAM). For the biggest models, the GPT-J-6B and GPT2-XL virtual machine parameters have been increased to 16vCPUs, 150GB of RAM, and 2x NVIDIA Tesla T4. Based on perplexity metrics, the best-rated model, in this case, is GPT. Meanwhile, considering only the new generation models, the best result is observed with the GPT2-Medium model. However, 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 GPTNeo-1.3B have a pretty similar RAM usage. The interesting fact is that the GPT model uses more RAM compared to GPT2 and DistilGPT2.

Data and Methods



Models used in this research: OpenAI Ada, Babbage, Curie, GPT, DistilGPT2, GPT2, GPTNeo, XLNet.

The models require high graphics card settings, and in this case, Nvidia T4 graphics cards are used. For this reason, and also to speed up the calculations, the Python Facebook DeepSpeed library was used. DeepSpeed is a deep learning optimization library to provide an easy and efficient use of distributed computing in model training. DeepSpeed uses ZeRO (Zero Redundancy Optimizer) optimization strategies / steps. 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	Peak RAM/VRAM usage (GB)	Fine-tuning time (hh:mm)	Perplexity	Avg. Score	Std. Deviation
Ada	-	-	-	41,81	10,05
Babbage	-	-	-	37,86	9,35
Curie	-	-	-	35,92	8,07
GPT	10.6/15	1:53	2,46	38,88	10,76
DistilGPT2	9.5/15	0:26	9,49	39,67	10,61
GPT2	10.5/14.5	0:46	10,26	43,25	11,42
GPT2-Medium	14.7/14.8	2:42	8,18	42,23	12,45
GPT2-Large	22.7/14.9	7:27	10,86	45,66	11,08
GPT2-XL	34.8/14.9	25:44	17,62	42,71	11,41
GPTNeo-125M	10.6/15	1:03	9,12	44,66	10,75
GPTNeo-1.3B	31.7/14.8	10:17	36,37	46,6	11,36
GPTNeo-2.7B	49/12.7	34:05	41,62	44,93	9,81
GPT-J-6B	101/15	72:25	37,08	-	-
XLNetBase	10.6/15	3:51	-	-	-
XLNetLarge	15.2/15	11:11	-	-	-

Conclusion

This article reviewed the architecture of transformers and the main models of transformers that can currently be used. Information is also provided on how model training should be performed in the case of models that are particularly large. And the idea of adapting natural language generation to business name generation is presented and is being further developed. Looking at the results of the study, it can be seen that when evaluating only the perplexity metric, it does not always show the best model. A particularly important assessment method for assessing natural language generation is human assessment, so a consumer survey was conducted in this study. 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. Against this background, the application 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, and XLNet models were not even suitable for this task.

https://www.researchgate.com/profile/Mantas-Lukauskas **https://www.linkedin.com/mlukauskas**