

Politechnika Krakowska
im. Tadeusza Kościuszki

Wydział Informatyki i Telekomunikacji

Kierunek: Informatyka

Przedmiot: Przetwarzanie Języka Naturalnego

Generowanie Tytułów Publikacji Naukowych

Autorzy: AG, MF, JG

semestr 7

grupa laboratoryjna 1

Kraków, grudzień 2024

1. Abstrakt

Współczesne osiągnięcia w dziedzinie sztucznej inteligencji w istotny sposób zmieniają podejście do automatyzacji procesów twórczych, w tym także generowania tekstów naukowych. Niniejszy raport przedstawia zastosowanie modelu językowego GPT-2 do generowania tytułów publikacji naukowych. Projekt koncentruje się na procesie trenowania modelu na specjalistycznym zbiorze danych oraz na analizie wyników uzyskanych podczas generowania tytułów. Dzięki wykorzystaniu zaawansowanych algorytmów językowych możliwe jest automatyczne tworzenie tytułów spełniających standardy akademickie. Wzorem dla tego projektu był istniejący projekt dostępny na GitHubie, stworzony przez Chandana Singha, który został zaadaptowany i rozwinięty na potrzeby naszego projektu

2. Wstęp

2.1 Cel

Celem projektu jest opracowanie narzędzia umożliwiającego automatyczne generowanie tytułów publikacji naukowych na podstawie wybranego zbioru danych. System powinien tworzyć tytuły zgodne ze standardami akademickimi, jednocześnie oszczędzając czas i wspierając proces twórczy badaczy.

Dodatkowo projekt ma na celu zbadanie, w jakim stopniu model GPT-2 jest w stanie generować tytuły o wysokiej jakości na podstawie specjalistycznych danych. Ważnym aspektem jest zapewnienie, aby generowane tytuły były nie tylko gramatycznie poprawne, ale także zgodne z konwencjami publikacji naukowych.

2.2 Zakres

Projekt obejmuje pozyskanie danych, ich przygotowanie oraz przeprowadzenie procesu trenowania modelu. Kolejnym krokiem jest wygenerowanie przykładowych tytułów oraz przeprowadzenie oceny jakości wyników. Szczególną uwagę zwrócono na eliminację błędów językowych oraz zapewnienie zgodności z konwencjami publikacji naukowych.

2.3 Metodyka

W projekcie wykorzystano model GPT-2, który został dostrojony (fine-tuned) na zbiorze danych arXiv (quant-ph) zawierającym tytuły z zakresu fizyki kwantowej, co umożliwiło trenowanie modelu pod kątem naukowego stylu. Proces dostrajania obejmował wielokrotne iteracje treningowe, optymalizację hiperparametrów oraz analizę wygenerowanych wyników. Model został dostrojony za pomocą techniki fine-tuning, co umożliwiło jego adaptację do specyficznego stylu akademickiego.

Część Teoretyczna

3.1 Modele Językowe i GPT-2

Modele językowe są systemami zdolnymi do przetwarzania i generowania tekstów w sposób zbliżony do ludzkiego. Są one trenowane na ogromnych zbiorach tekstów, dzięki czemu uczą się rozpoznawać wzorce i zależności językowe. GPT-2 (Generative Pre-trained Transformer 2) jest jednym z najbardziej zaawansowanych modeli językowych, który bazuje na architekturze transformera.

Architektura Modelu GPT-2

Model GPT-2 składa się z wielu warstw transformera, które pozwalają na efektywne przetwarzanie sekwencji tekstu. Transformer składa się z mechanizmów samoatencji (self-attention), które pozwalają modelowi analizować całe zdania, uwzględniając kontekst każdego słowa. Zastosowanie warstw liniowych oraz mechanizmów samoatencji pozwala na tworzenie tekstów zgodnych z kontekstem.

Proces Fine-Tuning

Fine-tuning to proces dostosowywania modelu do specyficznego zbioru danych, poprzez dodatkowe trenowanie na nowych danych wejściowych. Dzięki temu model może generować teksty bardziej zgodne z wymaganiami konkretnego zadania, takiego jak generowanie tytułów publikacji naukowych. W projekcie użyto tytułów publikacji z arXiv (quant-ph), co umożliwiło skoncentrowanie modelu na specyficznym stylu językowym.

Zasady Działania Modeli Generatywnych

Modele językowe generują tekst na podstawie sekwencji tokenów, które zostały już wygenerowane lub podane jako dane wejściowe. Obliczają prawdopodobieństwo wystąpienia kolejnego tokenu, biorąc pod uwagę kontekst całej wcześniejszej sekwencji. To podejście umożliwia tworzenie spójnych i logicznych tekstów, które są zgodne ze strukturą i stylem językowym.

Część Praktyczna

4. Realizacja Projektu

4.1 Wybór Zbioru Danych

Dane do trenowania modelu zostały pozyskane z bazy arXiv (quant-ph). Zbiór ten wybrano ze względu na jego dostępność, dużą liczbę tytułów oraz

specjalistyczny język naukowy. Zbiór danych składał się z tysięcy tytułów, co zapewniło wystarczającą bazę do treningu.

4.2 Proces Treningu

W celu dostosowania modelu GPT-2 przeprowadzono proces fine-tuningu na tytułach publikacji naukowych z dziedziny fizyki kwantowej. Dane zostały pobrane z serwisu arXiv przy użyciu biblioteki arxivscraper, a następnie zapisane w pliku titles_ref.csv. Proces treningu obejmował wczytanie danych oraz dostosowanie modelu przy użyciu biblioteki gpt_2_simple.

Model GPT-2 został dostrojony poprzez wielokrotne iteracje treningowe z konfiguracją parametrów, takich jak liczba epok, rozmiar wsadu oraz współczynnik uczenia. W trakcie treningu wyniki były monitorowane, a parametry dostosowywano w celu uzyskania najlepszej jakości generowanych tytułów. Przykłady wygenerowanych tekstów były analizowane pod kątem poprawności językowej oraz zgodności ze stylem akademickim.

W celu dostosowania modelu GPT-2 przeprowadzono proces fine-tuningu na tytułach publikacji naukowych, wykorzystując plik titles_ref.csv. Proces treningowy trwał 1000 kroków z zapisami pośrednimi i generowaniem próbek co 25 kroków.

```

1 import gpt_2_simple as gpt2
2
3 model_name = "124M"
4 gpt2.download_gpt2(model_name=model_name)
5
6 sess = gpt2.start_tf_sess()
7 gpt2.finetune(sess,
8               'titles_ref.csv',
9               model_name=model_name,
10              steps=1000,
11              save_every=200,
12              sample_every=25)
13
14 gpt2.generate(sess)
15

```

Do celów treningowych pobrano tytuły publikacji z arXiv za pomocą biblioteki arxivscraper. Dane te zostały zapisane w pliku titles_ref.csv jako zestaw treningowy.

```

1 import arxivscraper as ax
2 import numpy as np
3
4 scraper = ax.Scraper(category='physics', date_from='2019-05-01',
5                      date_until='2019-07-03', t=10,
6                      filters={'categories':['quant-ph']})
7
8 output = scraper.scrape()
9
10 titles = [' '.join(o['title'].split()) for o in output]
11 np.savetxt('titles_ref.csv', np.array(titles), fmt='%s')
12

```

Po zakończeniu treningu modelu przeprowadzono generowanie tytułów, zaczynając od prefiksu quantum. Wygenerowane tytuły były automatycznie formatowane i poddawane analizie pod kątem zgodności ze stylem naukowym.

```

1  import gpt_2_simple as gpt2
2  sess = gpt2.start_tf_sess()
3  gpt2.load_gpt2(sess)
4
5  prefix = 'quantum'
6  text = gpt2.generate(sess,
7                      length=40,
8                      temperature=0.7,
9                      prefix=prefix,
10                     nsamples=1,
11                     batch_size=1,
12                     return_as_list=True
13                )
14
15
16  t = text[0].title()
17  t = t.replace('<|Startoftext|>', '').replace('\n', '')
18  t = t[:t.index('<|Endoftext|>')]
19  print(t)
20

```

4.3 Przykłady Wyników

Przedstawiono wygenerowane tytuły, które były spójne gramatycznie i zgodne z tematyką fizyki kwantowej. Część tytułów wymagała jednak dalszej obróbki. Analiza wyników wskazała na konieczność dodatkowego dostrojenia modelu, szczególnie pod kątem bardziej skomplikowanych struktur językowych. Niektóre tytuły były zaskakująco kreatywne, jednak część z nich wymagała ręcznej korekty w celu lepszego dopasowania do standardów akademickich.

```

<|startoftext|>quantum interferometry with a four-photon entropic gap of langevin box models<|endoftext|>
<|startoftext|>simulation complexity and the nature of quantum mechanics<|endoftext|>
<|startoftext|>photon entanglement in long-range physical interactions<|endoftext|>
<|startoftext|>empirical determination of the exact quantum state signature of magnetic field<|endoftext|>
<|startoftext|>two-dimensional quantum dynamics test material: a volume-unconstrained model<|endoftext|>
<|startoftext|>anomalous quantum operations<|endoftext|>
<|startoftext|>strong interactions between superconducting qubits<|endoftext|>

<|startoftext|>characterizing nonlocal dispersion compensation in deployed telecommunications fiber<|endoftext|>
<|startoftext|>tree tensor networks for generative modeling<|endoftext|>
<|startoftext|>on 2-form gauge models of topological phases<|endoftext|>
<|startoftext|>a coherent superposition of feshbach dimers and efimov trimers<|endoftext|>
<|startoftext|>edge mode locality in perturbed symmetry protected topological order<|endoftext|>
<|startoftext|>quantum acousto-optic control of light-matter interactions in nanophotonic networks<|endoftext|>

```

5. Podsumowanie

Wpływ danych: Jakość danych wejściowych ma kluczowe znaczenie dla wyników modelu. Precyzyjnie wyselekcjonowany zbiór danych z arXiv pozwolił na uzyskanie spójnych tytułów.

Ograniczenia: Niektóre wygenerowane tytuły wymagały dalszej obróbki lub były zbyt ogólne, co wskazuje na ograniczenia modelu w zakresie generowania bardziej szczegółowych treści.

6. Bibliografia

- [GPT Paper Title Generator - GitHub](#)
- [arXiv Quant-ph Archive](#)
- <https://huggingface.co/docs/transformers/index>