Федеральное государственное автономное образовательное учреждение

высшего образования

«Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Фундаментальная и прикладная лингвистика

бакалавриат

профиль: компьютерная лингвистика

**О Т Ч Е Т**

**по проекту**

*Использование модели ruGPT-3 для решения задач NLP.*

Выполнил студент гр. 19ФПЛ-2

Шувар Алена Олеговна

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

*(подпись)*

**Проверил:**

*приглашенный специалист, Петрова Полина Дмитриевна*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(подпись)*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*(дата)*

**Описание цели проекта**

Русскоязычная модель GPT-3 (ruGPT-3) умеет продолжать тексты на русском языке, основываясь на шаблоне, который передает в модель человек – так называемом промпте, или “затравке”, например, при подаче модели текста “*ярмарка - женский род, конь - мужской род, солнце - средний род, вода -* ”, модель дополняет его следующим образом:

“*RuGPT3: ярмарка - женский род,*

*конь - мужской род,*

*солнце - средний род,*

*вода - женский род,*

*железо - этноним "скифы" - женского рода.*”

Проект предполагает изучение возможностей модели ruGPT-3 для решения различных задач NLP. Главной целью проекта являются анализ и составление подходящих вводных данных для моделей.

**Описание поставленной задачи в рамках групповой работы**

Мною была выбрана задача классификации текстов, в рамках которой необходимо придумать такие шаблоны для моделей ruGPT-3 различного размера, чтобы они могли корректно отнести тот или иной текст к определенному классу, например, при подаче модели текста новости “*Цена российской нефти для налогов осталась ниже $50 второй месяц подряд.*” модель смогла определить его в категорию “*Экономика*”.

**Подход к составлению шаблонов**

Поскольку модели необходимо классифицировать текст, кажется целесообразным не подвергать изменениям сам текст, а только менять форму его подачи – т. е. формулировку “задания”, которое будет прописываться для модели.

Были использованы следующие шаблоны:

1. «Текст новости. Категория: »;
2. «Есть три категории новостей: категория\_1 … категория\_n

Текст новости\_1. Категория: категория\_1

…

Текст новости\_n. Категория: »;

1. «Соотнесите новость с категорией из списка: категория\_1 … категория\_n

Текст новости\_1. Категория: категория\_1

…

Текст новости\_n. Категория: »

1. «Классифицируй текст. Классы: категория\_1 … категория\_n

Текст новости\_1. Категория - категория\_1

…

Текст новости\_n. Категория - »

1. «Классифицируй текст и напиши только одно слово. Классы: категория\_1 … категория\_n

Текст новости\_1. Категория - категория\_1

…

Текст новости\_n. Категория - »

**Описание использованных моделей**

В работе было использовано 3 модели ruGPT3, которые обучались с длиной контекста 1024, на корпусе из 80 млрд слов, 3 эпохи, затем были дообучены 1 эпоху с длиной контекста 2048:

1. ruGPT3Small: 125 миллионов параметров;
2. ruGPT3Medium: 350 миллионов параметров;
3. ruGPT3Large: 760 миллионов параметров.

Данные модели были реализованы в Google Colab со следующими гиперпараметрами:

1. do\_sample=True, max\_length=50, repetition\_penalty=5.0, top\_k=5, top\_p=0.95, temperature=1, num\_beams=None, no\_repeat\_ngram\_size=3;
2. early\_stopping=True, do\_sample=True, max\_length=250, repetition\_penalty=3.5, top\_k=5, top\_p=0.80, temperature=0.7, num\_beams=None, no\_repeat\_ngram\_size=3;
3. do\_sample=True, max\_length=212, repetition\_penalty=3.5, top\_k=10, top\_p=0.85, temperature=0.6, num\_beams=None, no\_repeat\_ngram\_size=3;
4. do\_sample=True, max\_length=212, repetition\_penalty=2.0, top\_k=3, top\_p=0.95, temperature=0.2, num\_beams=None, no\_repeat\_ngram\_size=3;
5. do\_sample=True, max\_length=230, repetition\_penalty=2.0, top\_k=5, top\_p=0.90, temperature=0.5, num\_beams=None, no\_repeat\_ngram\_size=3;
6. do\_sample=True, max\_length=50, repetition\_penalty=2.0, top\_k=5, top\_p=0.90, temperature=0.4, num\_beams=None, no\_repeat\_ngram\_size=3.

**Результаты и выводы**

В результате работы наиболее неэффективной оказалась маленькая модель (ruGPT3Small) – она практически не давала верных результатов, в связи с чем можно говорить о целесообразности использования ruGPT3Medium и ruGPT3Large для задачи классификации текста, при этом средняя модель в большинстве случаев справлялась так же хорошо, как и большая, что позволяет сэкономить вычислительные ресурсы при необходимости использования такого метода классификации текста.

В качестве “затравки” самым продуктивным оказался метод с текстом, содержащим императив глагола с семантикой указания на необходимое действие без каких-либо уточнений по размеру генерируемого текста (шаблон 4). Стоит отметить, что без передачи моделям конечного списка классов, она довольно хорошо классифицирует тексты сами.

Для контроля длины генерируемого текста, к сожалению, невозможно использовать шаблон типа “Напиши только одно слово”, поскольку эта “инструкция” не только не приводит к результату генерации 1 слова, но и может помешать моделям классифицировать текст с опорой на переданный ей список желаемых категорий текста. Чтобы ограничить длину генерируемой последовательности, необходимо вручную выставлять значение параметра max\_len, которое может варьироваться от модели к модели.

Кроме того, приведение нескольких примеров для одной категории “облегчает” модели классификацию, так, модель лучше предсказывает класс для текста, когда в качестве примеров она видела не 3 примера (реализации для затравок 1–2), а не менее 7 (затравки 3–5).

Средняя и большая модели могут понимать тексты с лексикой, которая не относится к литературному языку и может быть понятна не всем членам общества (сленг, профессионализмы, англицизмы и пр.), однако при этом возможна потеря связи с заданием и классификация не по заданным категориям.

Ссылка на ноутбук: <https://colab.research.google.com/drive/1jzHgTtaverRZmyOckvODo6wN9n_R_zjE?usp=sharing>