Силлабо-тоническая трансформерная языковая модель

для генерации стихов

Козиев Илья

Решаемая задача

Генерация стихотворных 4-строчников по задаваемой теме с соблюдением принятых в русской поэзии правил рифмования и метрики:

Я оставляю брошенные фразы
Иного смеха, слабости и слёз
Я превращаюсь в голубые стразы
Кружась ветвями молодых берёз

Почему не ruGPT и ruT5?

Обычная пословная GPT (e.g. rugpt) **плохо** генерирует стихи, потому что она не знает про **фонетику** русского языка. В ходе файнтюна такая gpt-модель может запомнить некоторое количество пар рифмуемых слов, но для стихов этого мало :(

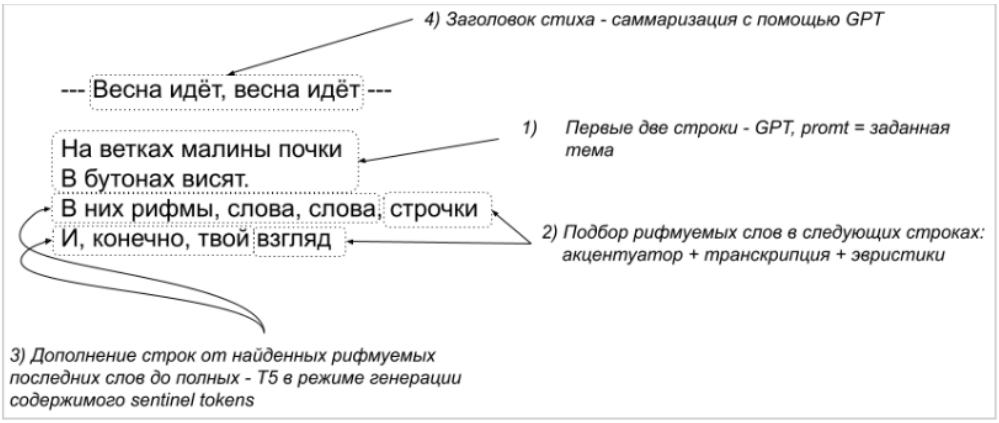
Иногда отфайнтюненая rugpt выдает нормальные стихи:

```
А в небе звёздочка горит,
Сияет так красиво!
И на ветвях берёз шумит
Своей листвой игриво
```

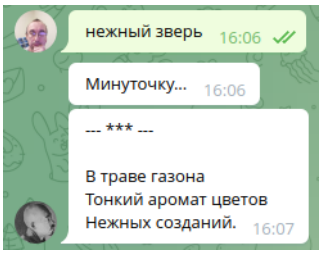
Хотя обычно получается “верлибр”:

```
Я знаю, ты не станешь смеяться
В этот день над моими стихами.
Только белая берёза проснётся
Под моим окном, и заплачет с нами
```

Можно ввести рифмовку принудительно, подбирая окончания строк 3 и 4 моделью рифмовки. Начала строк 3 и 4 - генерировать с помощью отфайнтюненой **ruT5** и **sentinel tokens**:



В случаях, когда рифмовка не обязательна, rugpt **вполне уместна**: см. генератор “хайку” https://t.me/haiku_guru_bot



Архитектура решения

Предлагаемое решение основано на **силлабо-тонической** языковой модели. При обучении эта модель в явном виде знакомится с фонетикой русской речи – делением на **слоги** и правилами чередования **ударных и безударных** слогов. Как и обычная GPT-модель, СТЯМ также учится **грамматике** русского языка.

В исходном тексте, на котором учится модель, размечаются **ударения**, слова разбиваются на **слоги** и цепочки слогов разворачиваются **справа налево** для каждой строки.

Исходный текст (автор - Ия Кузько):

```
И в нарастающей воде
Призывный стон из глубины
А дале гул звучит везде
То первый вздох и глас весны
```

После разметки, разбивки на слоги и “арабизации”:

```
де́ во | щей ю ста́ ра на | в | И <nl>
ны́ би глу | из | сто́н | вный зы́ При <nl>
зде́ ве | чи́т зву | гу́л | ле да́ | А <nl>
сны́ ве | гла́с | и | вздо́х | вый пе́р | То <nl>
```

Обучение состоит из претрейна на прозе и файнтюна на стихах.

Датасет для претрейна

Объем ~ **5 Гб прозы**

Примерно с 1 Гб русского текста модель начинает демонстрировать владение русской грамматикой, на 5 Гб количество ошибок падает до терпимого. Примеры ошибок:

```
Не заставляй меня стыдиться
маньяк любовных, нежных слов
Я так боюсь в тебе влюбиться
ах, как боюсь я дураков

Книги плотно прикрывая
Не словами в темноте
Я живу, не понимая
Отдаваюсь этой мечте

Я не могу Тобою упиваться
Мою любовь к тебе я прививать
Я без Тебя хочу, с Тобой остаться
А ты мне наше сердце отдавать
```

Разметка датасета – автоматическая, pos tagger Stanza и контексто-чувствительный ударятор, умеющий обрабатывать **омографы**.

Скорость разметки – около 0.4 Гб/сутки

Претрейн

Архитектура – **GPT**, кол-во весов примерно rugpt-small

Обучение **с нуля**

Свой токенизатор, ~25,000 токенов

Обучение – примерно 15 часов на 4 GPU, 1 эпоха

Файнтюнинг

Сырье – 4+ млн наскрапленных стихов, в основном **плохого** качества.

Свой поэтический транскриптор для расстановки ударений, учитывающий неударность клитик, опциональность ударений на местоимениях и т.д.

Из сырья **автоматически** отбираются стихи с хорошей рифмой, качественной метрикой, без грамматических ошибок, ~**400 тыс. четырехстрочников**.

Тексты 4-строчников подвергаются синтаксическому разбору, извлекаются именные группы, они становятся затравками для сэмплов:

```
буря ==> Буря мглою небо кроет …
```

Обучение: ~1.5 часа на 1 GPU, 1 эпоха

Инференс

Модель генерирует варианты стихов по заданной теме.

Используется top-p + top-k сэмплинг. Код сэмплинга из transformers модифицирован так, чтобы подавлять варианты с повтором рифмованного слова.

На CPU эта модель выдает 10 вариантов за 8-10 секунд.

Процент выхода годных стихов достаточно большой.

С незнакомыми словами и словосочетаниями модель **импровизирует**.

Пример для темы “субботняя коза”:

```
Люблю всегда и каждому в глаза
О, боже, как люблю я наши годы
Ах, Нина, ты – субботняя коза
Погода, счастье, неудачи и невзгоды
```

Примеры

Доступен бот в телеграмме @verslibre_bot (https://t.me/verslibre_bot)

```
          * * *
Я оставляю брошенные фразы
Иного смеха, слабости и слёз
Я превращаюсь в голубые стразы
Кружась ветвями молодых берёз


          * * *
В зеркале красивых снов
Образ мы читаем смело
Образ в зеркале ветров
Образ, что скрывает тело


          * * *
Открыв глаза, увижу сон
И лишь однажды я узнаю
Я боль, когда я вижу сон
Вдыхаю, выдыхаю, таю


            * * *
Найдётся счастье в неизбежности
Ты не из тех, кому не жить
Коронавирус страсти, нежности
Чтоб ненавидеть и любить
```

Другие способы генерации

Можно файнтюнить модель просто на стихах, чтобы она дописывала первые строки:

```
       * * *
Мчатся тучи, вьются тучи
Опускается рассвет
Мысли пламенные жгучи
И весны тревожный след

            * * *
Я достаю из широких штанин
Из благородной затейливой пыли
И я всё такой же, совсем как один
Я рад, что вы обо мне позабыли

            * * *
Белая берёза под моим окном
Золотистый свет её струится
Хочется забыться вечным сном
И отчаяться понять всё то, что снится

        * * *
Белая береза
Под моим окном
Принакрылась снегом
И забылась сном
```

Спасибо за внимание!