Может ли машинное обучение быть исследовательским инструментом?

Борис Орехов (НИУ ВШЭ, ИРЛИ РАН, МГУ)

В докладе не будет темы LLM

Машинное обучение

Машинное обучение — это

- Концепция, в которой предполагается создание некоторой модели, умеющей принимать решения,
- но при этом алгоритм сам «обозревает» материал (обучающие данные), «обучается»
- и статистически выводит закономерности
- распознавание паттернов посредством статистической индукции

Использование машинного обучения в исследованиях

Подготовка данных для исследования.

Например, разметка корпусов

Мама	мыла	раму
v мама x СУЩ, од, жр, ед, им	v мыло x СУЩ, неод, ср, ед, рд	v рам x СУЩ, неод, мр, гео, ед, дт
	v мыло x СУЩ, неод, ср, мн, им	v рама x СУЩ, неод, жр, ед, вн
	v мыло x СУЩ, неод, ср, мн, вн	
	v мыть x ГЛ, несов, перех, жр, ед, прош, изъяв	

Можем ли мы использовать машинное обучение

как исследовательский инструмент?

Какое знание ищут ученые?

Среди прочего — скрытые паттерны

Например, лингвисты ищут ключ к распределению фонем в слове, фонетических явлений в речи, лексем в текстах

Распознавание паттернов — это буквально определение машинного обучения

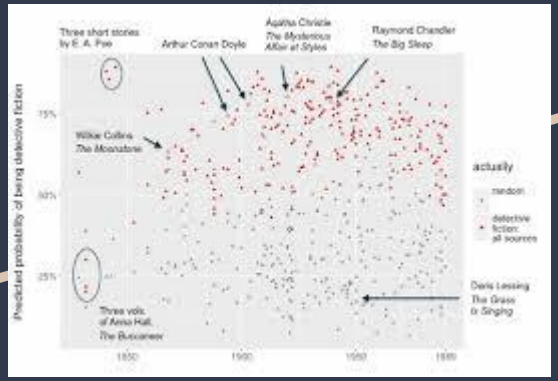
Пример без машинного обучения

Орехов Б. В. Нестиховедческий ритм в романе Н. Г. Чернышевского «Что делать?» // Цифровые гуманитарные исследования. — 2024. — № 1. — С. 34–44.

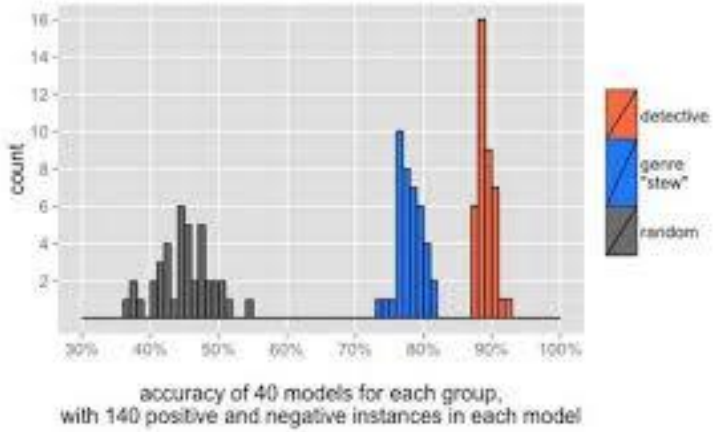
Цифровые гуманитарные исследования

Ted Underwood

The Life Cycles of Genres



Вопрос о границе жанров решается с помощью тренировки модели



Стиховедческое исследование: размеры силлаботоники

Что нужно, чтобы машина приняла решение об определении размера?

то же, что нужно и человеку

Ключевое понятие силлаботоники — это ударение (ударный слог).

Этот тезис выглядит как трюизм.

Сама идея отрицания ключевой роли ударения для определения размера выглядит абсурдно.

Стихи, порожденные рекуррентной символьной сетью

И так же всё теперь слова.

Проснись ль скорей и воздух дал.

Но вот к конце весенних дней

И облачное на челе,

Спешит полутвердил я в небе,

И не сказал бы все врага,

И от меня в нем был лишь славой.

Сеть обучена на метрически гомогенных данных.

Вопрос: наличие скрытой системности в данных

Обучающая выборка не содержит данных об ударении (!)

Генеративная модель «угадывает» размер без ударения.

Нейросети этой архитектуры умеют не только порождать текст, но и классифицировать текст.

Нужно попробовать в «обратную» сторону: угадает ли размер в сходных условиях классифицирующая модель?

Если **нет**, то наше представление о размере и ударении было верным.

Если **да**, то ударение не обязательно для определения размера, что мы раньше назвали абсурдным.

Машинное обучение здесь не прикладной, а исследовательский инструмент

Это не прикладной, а исследовательский вопрос.

Машинное обучение позволяет доказать наличие системы там, где мы не знаем, есть ли система.

Машинное обучение не позволяет опровергнуть наличие системы.

Бинарная классификация

Сеть может «понять», что на размер работает количество гласных.

Поэтому проверим, умеет ли сеть различать размеры с одинаковым числом слогов.

Сеть может понять, что ударение — это «слово» (группа графем между пробелами).

Поэтому проверим, умеет ли сеть различать размеры с одинаковым числом иктов.

8 слогов

Я4м

Х4ж

~~Д3ж~~

Аф3м

~~Ан2д~~

Обучающие данные

100 000 строк Я4м

100 000 строк Х4ж

Пример:

"Я4м": "А сердцем больше щедр и благ", "Поверхность вод позолотит", "К ее пленительным устам"…

"Х4ж": "«Ничего... Жену спровадил", "Нам лишь чудо путь укажет", "О бананах долгоплодых"…

Каждой строке соответствует метка размера.

Модель обучается тому, что такая последовательность букв соответствует такому размеру.

Нет данных об ударении! Только буквы.

Оценка модели

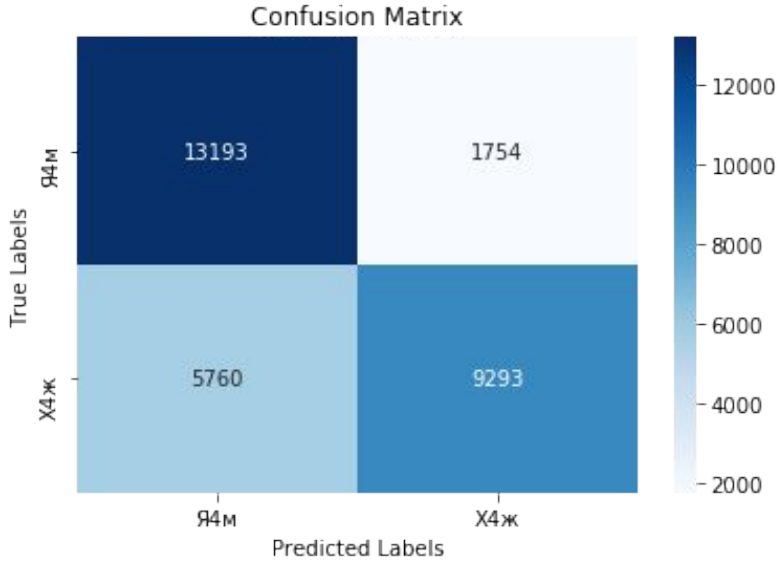
0,7495

	точность	полнота	f1-мера	всего
Я4м	0.6961	0.8827	0.7783	14947
Х4ж	0.8412	0.6174	0.7121	15053

Обучили модель. Теперь протестируем ее.

Данные для оценивания:

30 000 строк, которые сеть не видела в процессе обучения.



Оценка модели

Это значит, что в 75 % случаев размер определяется правильно.

Если бы речь шла о случайности, то значение было бы 50 %.

В предельной формулировке это значит, что ударения не нужны для определения силлабо-тонического размера.

Нужны ли ударения?

Конечно, ударения нужны — для точного определения размера.

Но эксперимент показывает, что размер «звучит» и на других уровнях текста, кроме просодического.

Это удивительный факт.

До эксперимента он выглядел абсурдно.

Почему так вышло?

Модель ориентируется только на распределения букв (не слов!).

Самыми частотными сочетаниями букв в тексте наряду со служебными словами являются морфологические форманты (суффикс+окончание).

Форманты различают части речи.

Благодаря лингвистике стиха мы знаем, что распределение частей речи в строке не случайны.

Прилагательные предпочитают вторую половину строки, особенно третью стопу (...) Что касается наречий, то они предпочитают первую половину строки (Лингвистика стиха, с. 67)

По всей видимости, сеть выучивает тенденции в распределении морфологических форм в строке, характерном для конкретного размера.