

Стих как тип текста: проверка машинным обучением

Борис Орехов

НИУ Высшая школа экономики

nevmenandr@gmail.com

25 октября 2017

- 1 Стих как объект лингвистического исследования
- 2 Стих и проза
- 3 Машинное обучение
- 4 Эксперимент

- 1 Стих как объект лингвистического исследования
- 2 Стих и проза
- 3 Машинное обучение
- 4 Эксперимент

Правда ли, что это литературоведение?

- Это 5-я конференция по общему, скандинавскому и славянскому языкознанию.
- Доклад про стих, значит, что это литературоведение?
- *Herbelot A. The semantics of poetry: A distributional reading // Literary and Linguistic Computing. 2014.*
- *Обычный* язык и правила синтаксиса и семантики vs. язык поэзии.
- В поэтическом тексте происходят не совсем обычные языковые процессы. Они могут быть исследованы лингвистическими средствами.

Есть и лингвистика стиха



- Что изучают литературоведы?
- Является ли стиховедение литературоведческой дисциплиной?
- Квантитативный подход. Статьи ак. А. Н. Колмогорова.
- Является ли стиховедение лингвистической дисциплиной?

Что такое стих? Немного о терминах

- Стихи — это тексты, созданные по определённым правилам.
Исторически — не вполне понятно, зачем.
- Стих — это
 - ① одна строка (например, *стих 48*)
 - ② совокупность характеристик, релевантных для понимания, как устроена ритмическая организация речи
 - ③ **не** стихотворение
- Что за правила?

- 1 Стих как объект лингвистического исследования
- 2 Стих и проза**
- 3 Машинное обучение
- 4 Эксперимент

Поэзия и проза отличаются друг от друга

- по статусу (или имиджу)
- структурно

Расхожее представление состоит в том, что стихи труднее писать, следовательно, они являются более престижной формой речи (хотя хорошую прозу писать не легче).

Кроме того, стихотворная форма исторически связана с магическими практиками

Господин Журден

Г-н Журден. А теперь я должен открыть вам секрет. Я влюблен в одну великосветскую даму, и мне бы хотелось, чтобы вы помогли мне написать ей записочку, которую я собираюсь уронить к ее ногам.

<...>

Учитель философии. Вы хотите написать ей стихи?

Г-н Журден. Нет, нет, только не стихи.

Учитель философии. Вы предпочитаете прозу?

Г-н Журден. Нет, я не хочу ни прозы, ни стихов.


Учитель философии. Так нельзя: или то, или другое.

Г-н Журден. Почему?

Учитель философии. По той причине, сударь, что мы можем излагать свои мысли не иначе, как прозой или стихами.

Г-н Журден. Не иначе, как прозой или стихами?

Учитель философии. Не иначе, сударь. Все, что не проза, то стихи, а что не стихи, то проза. <...>

Г-н Журден. Честное слово, я и не подозревал, что вот уже более сорока лет говорю прозой. Большое вам спасибо, что сказали. 

Структурные отличия стихов от прозы. Тупиковые варианты

- ритм
- рифма
- цветы и птицы

А что такое ритм?

- ритм — последовательность повторяющихся или варьирующихся элементов



Ритм — универсальное понятие, он есть и в прозе

Все **счастливые семьи** похожи друг на друга, каждая **несчастливая семья** несчастлива по-своему. **Все смешалось** в доме Облонских.

Жена узнала, что **муж был в связи** с бывшею в их доме француженкою-гувернанткой, и объявила мужу, что не может жить с ним в одном доме. **Положение это продолжалось** уже третий день и мучительно чувствовалось и самими супругами, и всеми членами семьи, и домочадцами.

Стихотворный ритм?

Может быть, различению может служить «специальный»
традиционно понимаемый стихотворный ритм?

Она пришла с мороза,
Раскрасневшаяся,
Наполнила комнату
Ароматом воздуха и духов,
Звонким голосом
И совсем неуважительной к занятиям
Болтовней.
Она немедленно уронила на пол
Толстый том художественного журнала,
И сейчас же стало казаться,
Что в моей большой комнате
Очень мало места.
Все это было немножко досадно
И довольно нелепо.

Впрочем, она захотела,
Чтобы я читал ей вслух «Макбета».
Едва дойдя до пузырей земли,
О которых я не могу говорить без волнения,
Я заметил, что она тоже волнуется
И внимательно смотрит в окно.
Оказалось, что большой пестрый кот
С трудом лепится по краю крыши,
Подстерегая целующихся голубей.
Я рассердился больше всего на то,
Что целовались не мы, а голуби,
И что прошли времена Паоло и Франчески.
1908

In nova fert animus mutatas dicere formas
corpora; di, coeptis (nam vos mutastis et illas)
adspirate meis primaque ab origine mundi
ad mea perpetuum deducite tempora carmen.

Оно же:

Ныне хочу рассказать про тела, превращенные в формы
Новые. Боги, — ведь вы превращения эти вершили, —
Дайте ж замыслу ход и мою от начала вселенной
До наступивших времен непрерывную песнь доведите

Р. О. Якобсон:

В эпоху классицизма или романтизма репертуар поэтических тем был весьма ограничен. Мы все хорошо помним традиционный реквизит: месяц, озеро, соловей, скалы, розы, замок и т. д. и т. п. Даже сны романтика не выходили за рамки этого круга. <...> Особенно готические окна были в моде, и за ними обязательно светила луна. Сейчас для поэта одинаково поэтично любое окно, начиная с огромной зеркальной витрины универмага и кончая окошком деревенского трактира, густо засиженным мухами.

И через поэтические окна ныне можно увидеть всякое. Незвал писал об этом:

Посреди фразы меня вдруг ослепит сад
или нужник неважно что
я уже не различаю предметы по приписываемой им вами
прелести или
безобразию.

Знаете ли вы венгерский?

Hallgass, bújj el, s titkold, tagadd
érzéseid, álmaidat!

Mint fénylő csillagmiriád
szállhatnak a lelkedен át,
érkezve s tűnve, mint az éj:
csodáld őket és — ne beszélj!

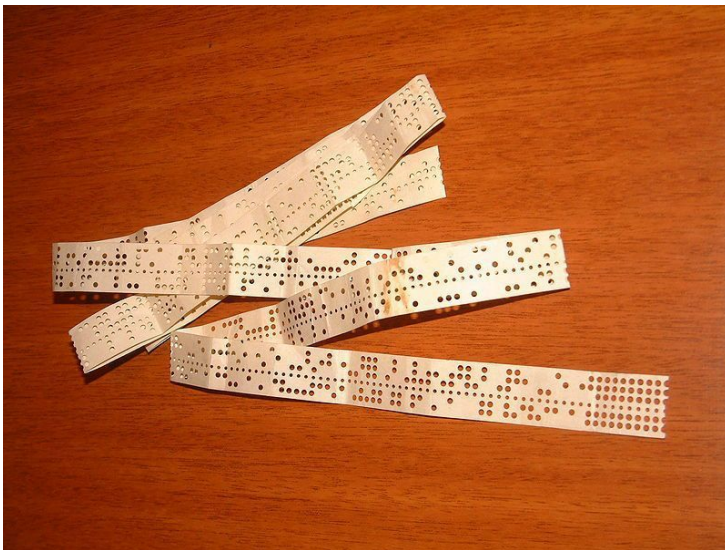
Szív hol s kinek nyílhatna meg?
Ki értheti az életed?
Ki érthetné, ki vagy, mi vagy?
Hazudik a kész gondolat!
Merítve sár a tiszta mély:
igyál belőle s — ne beszélj!

Это стихи или проза?

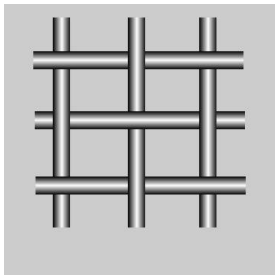
A Goldberg-variációk Bach egyetlen variációformában komponált műve, amelyet Hermann Karl von Keyserlingk gróf és csembalistája, Johann Gottlieb Goldberg számára írt 1741–42-ben. A mű átfogó, akadémikus alkotás, a barokk variációművészet mintapéldája. Minden rész önálló karakterrel, egyedi hangulattal bír, amelyek összhangját a tudatosan építkező szerkezet és az erőteljes, kánon alakzatokban rendszeresen visszatérő egyetlen közös basszustéma biztosítja.

A szokatlanul hosszú — 32 ütemes — főtéma harmóniameghatározó vázhangjai mind a harminc variációban változatlanok maradnak, míg az ívmotívum variált formákban ismétlődik meg. A műben háromféle variációtípus váltakozik: a fokozódó virtuozitással komponált futamokból, arpeggiókból álló „figuratív variációkat”, a különböző formák és műfajok stílusában — triószonáta, tánc, siciliano, fuga, szólóconcerto, ária, francia nyitány, quodlibet — felhangzó „karaktervariációk” követik, majd az ária basszustémája fölött egyre nagyobb belépési távolságokkal (a prímtől egészen a nonáig) megalkotott „kánonok” zárják.

Проза как телетайпная лента



Стихи как система парадигматических членений



Еще в полях белеет **снег**,
А воды уж весной **шумят** —
Бегут и будят сонный **брег**,
Бегут и блещут и **гласят**...

Стих — это система сквозных принудительных парадигматических членений, структурирующих дополнительное измерение текста

(М. И. Шапир)

Можно ли это определение формализовать?

- Формализация — это методологическая проверка ясности формулировки.
- Кажется, что нельзя, потому что не ясно, что создаёт парадигматические членения, а что нет.
- Другое определение:

«стих — это прежде всего речь, четко расчлененная на относительно короткие отрезки, соотносимые и соизмеримые между собой. Каждый из таких отрезков тоже называется “стихом” и на письме обычно выделяется в отдельную строку»

(М. Л. Гаспаров)

Соотносимость и соизмеримость

- Оба определения интуитивные, никто их не проверял.
- Соотносимость — это то же парадигматическое членение.
- Соизмеримость — это уже интереснее.
- Можно сказать, что это соизмеримость длин строк?
- Тогда можно проверить определение на прочность.
- Но в чём мерить строки?

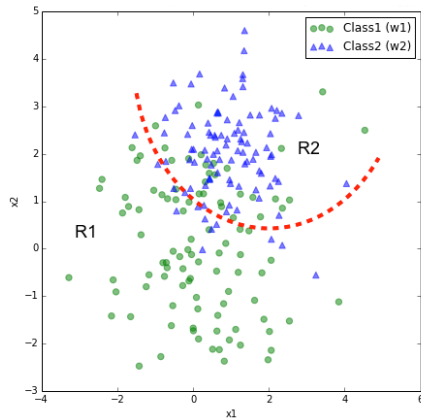
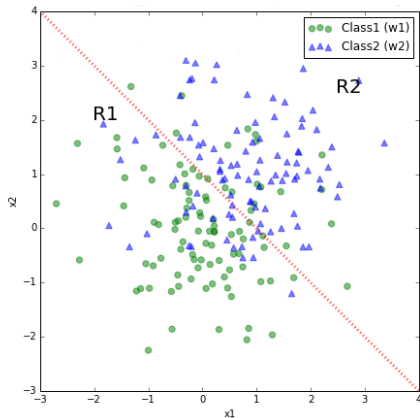
Содержание

- 1 Стих как объект лингвистического исследования
- 2 Стих и проза
- 3 Машинное обучение
- 4 Эксперимент

Что такое машинное обучение?

- Это набор алгоритмов, которые, «смотря» на исходные (обучающие) данные, пытаются вывести функцию, которая бы хорошо описывала некоторые числовые параметры этих данных.
- Например, зависимость значения одного параметра от другого.
- Может применяться для предсказания событий или распознавания (классификации) объектов.
- Эта функция — субоптимальное решение, то есть такое, которое является приемлемым там, где нахождение оптимального решения принципиально невозможно.

Как это выглядит?



Что мы будем предсказывать?

- Ничего.
- Но мы можем проверить, действительно ли возможно построить такую функцию, пользуясь нашей трактовкой определения Гаспарова?
- Если определение верно, то мы найдём функцию, которая будет хорошо предсказывать, стихи перед нами или проза, когда на вход функции будет подаваться только информация о соизмеримости строк.
- Если теоретики ошибаются, то мы такой функции не найдём и машина всё время будет ошибаться в определении типа текста.
- Обычно поступают по-другому!

- 1 Стих как объект лингвистического исследования
- 2 Стих и проза
- 3 Машинное обучение
- 4 Эксперимент**

- Мы вручную разметили (стих или проза) 2300 текстов разного объема, взятых из
 - stihi.ru,
 - «Арион»,
 - «Критическая масса»,
 - «Логос»,
 - «Новое литературное обозрение»,
 - «Неприкосновенный запас»,
 - «Октябрь»,
 - «Вопросы литературы».
- Общий объем выборки составил 117657 прозаических строк, 62196 стихотворных строк,
- в словоупотреблениях: 5954231 токен и 298103 токена.

Что сделано?

- Все тренировочные данные были преобразованы в табличный формат
- каждой строке соответствовало вхождение отдельной строки текста
- колонке — один из признаков этой строки.
- Признаками при этом выступали длина описываемой строки и длины соседних с ней строк (4 строки до 4 строки после).
- То есть соизмеримость текстовых отрезков, если она действительно служит одним из ключевых свойств стихотворной речи, должна распознаваться машинными алгоритмами в процессе обучения

Пример разметки

<line="p"/>И действительно, Наталья Горбаневская начинала сотрудничать с организованным литературным андеграундом — самиздатскими журналами «Синтаксис» и «Феникс» — в качестве художника. Однако до этого она уже была хорошо известна как самостоятельный неофициальный поэт.

<line="p"/>Как уже отмечалось, первое стихотворение, которому Горбаневская позволила числиться «среди живых», относится к 1956 году.

<line="v"/>Данный мир

<line="v"/>удивительно плосок.

<line="v"/>Прочий

<line="v"/>заколотен наглухо.

<line="v"/>Не оставили даже щель между досок.

<line="v"/>Старались. Мастера.

Таблица: Таблица данных

N	n	n-1	n-2	n-3	status
43	10	11	0	10	v
44	11	10	11	0	v
12	792	593	635	515	p

Таблица: Результаты

Alg	Right	Acc	F-m
Naive	7643	0.446	0.527
Bayes	15846	0.924	0.955
NearestCentroid	16076	0.938	0.961
KNeighbors	15976	0.932	0.957
Random	16494	0.962	0.977
Forest	15393	0.898	0.934
LRegression	16524	0.964	0.978
DecisionTree			
StochasticGradientDescent			

Всё ли в порядке в этом эксперименте?

- На самом деле, нет.
- Дело в том, что в обучающие данные мы включили длину строки.
- Что если алгоритм выучивается определять стихи как короткие строки?
- Проведём эксперимент ещё раз, исключив данные о длине строки.
- Теперь — только разница.

Таблица: Только дельта

Alg	Right	Acc	F-m
NB	6131	0.358	0.411
NC	15821	0.923	0.954
KN	15895	0.927	0.954
RF	15895	0.927	0.954
LR	16522	0.964	0.978
DT	15275	0.891	0.929
SGD	16499	0.962	0.977

Где ошибается алгоритм?

vav12.html, 9

<line="p"/> Пастернаковское назидание по поводу архивов и рукописей Цветкову не за чем принимать к сведению: он - и здесь я не встречал ему равных - с великолепным равнодушием относится к уже написанному, более того, на том стоит:

<line="v"/> ...Помнишь, у тебя

<line="v"/> (у нас, пожалуй) был такой приём

5 72 84 196 5 5 6 5 v

stihi4.html, 1

stihi4.html, 3

<line="p"/>Руби Штейн

<line="p"/> Вы смеетесь надо мной, потому что отличаюсь от вас,

<line="p"/> а я смеюсь над вами, потому что вы не отличаетесь друг
от друга.

<line="p"/> (Михаил Булгаков)

14 17 17 17 5 11 17 17 p

5 19 22 22 16 22 22 22 p

16 11 3 6 6 6 6 4 p

- Определение устояло.
- Машинное обучение, применённое «наоборот», подтвердило значимость соизмеримости строк для стиха.