Б. В. Орехов

Национальный исследовательский университет «Высшая школа экономики»
(Россия, Москва)
nevmenandr@gmail.com

Соизмеримость стиховых сегментов: проверка машинным обучением

В статье классическая проблема соотношения стиха и прозы решается в практической плоскости с использованием алгоритмов машинного обучения. На обучающей выборке объемом 2300 текстов (117 657 прозаических строк и 62 196 стихотворных строк) были натренированы модели с использованием алгоритмов Naive Bayes, Random Forest и Support Vector Classification. Признаками в одном случае выступали длины строк (описываемой и соседних с описываемой в окне 4), а в другом — разница в длине строк-соседей, взятая по модулю. Тестовая выборка состояла из 100 текстов, объемом 17 146 строк. Выяснилось, что хуже всего работает классификатор, построенный на Naive Bayes, в то время как Random Forest способен правильно классифицировать 93 % строк. Support Vector Classification показал средние значения полноты и точности. Другая модель была построена на основном и поэтическом подкорпусах НКРЯ, и показала самые низкие результаты на предложенных тестах. Из проверки следует, что соизмеримость длин строк в слогах действительно является важным критерием выделения стихотворной строки. При этом длина строки в буквах не играет значимой роли.

Ключевые слова: стих, проза, машинное обучение, наивный байес, автоматическая обработка текста, поэтический корпус.

Проблема разграничения стиха и прозы на теоретическом уровне решается давно и по-разному. Не может быть сомнений в том, что сама оппозиция стиха и прозы возможна только при отрефлектированности понятия стиха, ведь проза, как известно, представляет собой немаркированный член этой оппозиции («немаркированный» в том смысле, который сформулирован для таких словоупотреблений В.А. Плунгяном: «более простой», «более распространенный», «базовый» [Плунгян 2011: 25]). Таким образом, для того, чтобы отделить стих от прозы, нужно предельно ясно представлять себе, что такое стих. Конвенциональное определение стиха мы возьмем у М.Л. Гаспарова (хотя истоки этого определения были выработаны в науке раньше): «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» [Гаспаров 2004: 7]. Несмотря на то, что стиховедение выделя-

ется среди других гуманитарных дисциплин своим стремлением к точным методам и эмпирической базе, есть основания полагать, что в основе этой дефиниции лежат общие соображения. Иными словами, нам неизвестны сделанные на широком материале исследования, которые подтверждали бы, что в отличие от прозы стиховые отрезки действительно соизмеримы. Само по себе такое положение вещей не удивительно: ученые сравнительно недавно получили в свое распоряжение обширные текстовые коллекции, которые могут служить материалом подобного рода экспериментов. Соответственно, и плохо поддающиеся ручному выполнению подсчеты стали производиться на компьютере только в последние годы (разумеется, компьютерные исследования текстов проводились уже несколько десятилетий назад, но они по необходимости были ограничены скромными объемами выборок).

Мы подошли к проверке цитированного определения на прочность со следующих позиций. Во-первых, мы попытаемся транслировать его теоретическую основу в численно измеримые категории. Во-вторых, используем практико-ориентированный подход; т. е. признаем определение прошедшим проверку в том случае, если оно сможет помочь автоматически определить стихотворные отрезки в тексте. Для этого мы применим алгоритмы широко используемого в современных инженерных и исследовательских задачах машинного обучения. Машинным обучением называется комплекс статистических решений, которые позволяют, основываясь на наборе признаков некоторых известных объектов, предсказывать статус других, неизвестных, объектов. Так как эти решения требуют объемных вычислений, обычно они производятся с помощью компьютера (и поэтому называются машинными). В нашем случае объектами будут выступать строки текста, а их признаками — разница в длине соседствующих строк, так как в качестве исходного допущения мы приняли, что постулируемая «соизмеримость» охватывает прежде всего длину строки в слогах. Именно число слогов, а не графем (букв) стало для нас базовой метрикой, так как именно слоги (их количество и качество) являются основой всех описанных систем стихосложения. Системы стихосложения, в которых ключевым признаком были бы буквы, нам неизвестны.

Отдельно отметим, что другое существующее определение стиха, несмотря на его теоретическую обоснованность, трудно конвертировать в практически измеримые параметры: «стих — это система сквозных принудительных парадигматических членений, структурирующих дополнительное измерение текста» [Шапир 2000: 83]. Спектр явлений,

в которых могут себя проявлять парадигматические членения, настолько широк, что определить границы дополнительного измерения текста не представляется возможным. Так что для описываемого исследования это определение приходится признать непригодным, что, однако, не означает его низкой оценки. Отметим лишь, что в рамках используемой научной конфигурации дефиниция М. И. Шапира остается непроверяемой.

В задачах машинного обучения (точнее — их разновидности под названием «обучение с учителем») исследователь должен заранее заготовить обучающую выборку, на которой можно было бы тренировать модель, то есть специальным образом статистически распределить значения признаков так, чтобы наборы этих значений соответствовали некоторому статусу объектов. В применении к нашей задаче речь идет о тренировочном наборе текстов, в котором каждой строке был бы присвоен статус («стих» или «проза») и подсчитана ее длина. Мы вручную разметили 2300 текстов разного объема, взятых с сайта stihi.ru и из периодических изданий «Арион», «Критическая масса», «Логос», «Новое литературное обозрение», «Неприкосновенный запас», «Октябрь», «Вопросы литературы». Общий объем выборки составил 117657 прозаических строк, 62196 стихотворных строк, им соответствуют следующие объемы в словоупотреблениях: 5954231 токен и 298103 токена. Все тренировочные данные были преобразованы в табличный формат, где каждой строке соответствовало вхождение отдельной строки текста (прозаического параграфа или стиха), а 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). За такой исследовательской стратегией стоит следующая идея: соизмеримость текстовых отрезков, если она действительно служит одним из ключевых свойств стихотворной речи, должна распознаваться машинными алгоритмами в процессе обучения и в дальнейшем служить хорошим основанием для предсказания того, со стихом или с прозаическим абзацем имеет дело машина при анализе незнакомого текста. Если же соизмеримость не позволит хорошо классифицировать строки на прозаические и стихотворные, значит, соизмеримость стихотворных строк следует признать не самым принципиальным параметром для определения стиха.

Следует сказать, что именно машинное обучение лучше всего подходит для выполнения такого рода задач, потому что как раз такого рода системы всегда предлагают субоптимальное решение проблемы, то

есть такое, которое является приемлемым там, где нахождение оптимального решения принципиально невозможно. В самом деле, вряд ли осуществимо создание строгой математической функции, которая бы с полной точностью описывала соизмеримость стиховых сегментов во всех случаях их вариаций в реальных текстах. В то же время нахождение субоптимального решения с точностью и полнотой выше некоторого порога позволит сказать, что соизмеримость стиховых сегментов действительно в существенном проценте случаев способна дать различить стих и прозу. Пример получившихся табличных данных в табл. 1.

Таблица 1

Пример табличных данных,
поданных на вход алгоритму машинного обучения

№ текста и строки	Длина строки n	Длина строки n-1	Длина строки n-2	Длина строки n-3	Длина строки n-4	Длина строки n+1	Длина строки n+2	Длина строки n+3	Длина строки n+4	Статус
256, 42	11	0	10	11	10	10	11	10	0	стих
256, 43	10	11	0	10	11	11	10	0	11	стих
256, 44	11	10	11	0	10	10	0	11	10	стих
256, 45	10	11	10	11	0	0	11	10	11	стих
1020, 12	792	593	635	515	111	0	5	167	816	проза

В предметной области машинного обучения задачи, подобные описанной, называются задачами классификации, то есть объекты классифицируются (распределяются по классам) в зависимости от своих признаков. Мы применили наиболее популярные алгоритмы классификации Naive Bayes, Random Forest и Support Vector Classification.

Для проверки работоспособности получившихся моделей была вручную создана тестовая выборка, состоящая из текстов с сайтов stihi.ru (не совпадающих с текстами, входящими в обучающую выборку) и vavilon.ru, общим объемом 17146 строк (из которых 3177 прозаических и 13 969 стихотворных). Результаты представлены в таблице 2. В ней для каждого алгоритма сообщается число правильных предсказаний типа строки в тестовой выборке, а также значения специальных метрик оценки алгоритма: accuracy (доля строк, в отношение которых классификатор принял правильное решение) и F-мера (одновременная характеристика точности и полноты в решениях классификатора).

Таблица 2

Результат проверки моделей, основанных на длинах строк

Алгоритм	Число правильных решений (из 17146)	Accuracy	F-мера
Naive Bayes	3806	0.221	0.205
Support Vector Classification	11802	0.688	0.763
Random Forest	15974	0.931	0.956

Результаты говорят о том, что Naive Bayes плохо справляется с предложенной задачей и демонстрирует низкую долю правильных решений. Это неудивительно: особенностью байесовского подхода к данным является представление о независимости признаков друг от друга, то есть лучше всего алгоритм должен был бы предсказывать класс у тех строк, длины соседей которых в точности повторяли бы значения обучающей выборки, что достаточно редкий случай, особенно для прозаических параграфов. Удовлетворительно показал себя Support Vector Classification, верно предсказавший класс более половины строк тестовой выборки. С лучшей стороны себя проявил алгоритм Random Forest, который выдал правильное решение в 93 % случаев.

На следующем шаге мы создали новые модели, в которых признаками стали не длины строк, а взятая по модулю разница в длине между описываемой строкой и её соседями. При этом сама длина описываемой строки была убрана из числа признаков. Такое представление данных в большей степени позволило сосредоточиться именно на соизмеримости строк в то время как предыдущий эксперимент выводил на первый план вторичный по отношению к соизмеримости параметр — собственно длины строк.

Таблица 3

Результат проверки моделей, основанных на разнице в длинах строк

Алгоритм	Число правильных решений (из 17146)	Accuracy	F-мера
Naive Bayes	7643	0.445	0.527
Support Vector Classification	12963	0.756	0.824
Random Forest	15906	0.927	0.954

Результаты проверки модели, построенной на разницах в длинах строк (табл. 3), оказались лучше почти для всех опробованных алгоритмов. Это тоже не должно удивлять, так как разница в длинах должна

давать более унифицированные значения и разброс их в модели должен оказаться меньше, что, в свою очередь, позволяет лучше классифицировать новые случаи из тестовой выборки. Иными словами, строки в обучающей выборке должны были оказаться в большей степени похожими по значениям параметров на строки в тестовой выборке.

Впрочем, этот вывод не касается показателей алгоритма Random Forest, который даже несколько ухудшил свое классификационное качество.

Следующей проверке подверглась модель, построенная на основном и поэтическом подкорпусах НКРЯ (общий объем: 5 144 306 строк), в качестве признаков были избраны взятые по модулю разницы в длинах соседних строк. Разметка производилась автоматически: было принято необходимое допущение, что в основном корпусе все тексты прозаические. На деле это не соответствует действительности, так что в обучающей выборке присутствовал неизбежный процент брака. К сожалению, нам не удалось за приемлемое время получить модель с использованием алгоритма Support Vector Classification, который слишком медленно работал на объемных обучающих данных, так что дальнейшие результаты в таблице 4 приведены только для Naive Bayes и Random Forest.

Таблица 4

Результат проверки моделей, построенных на подкорпусах НКРЯ

Алгоритм	Число правильных решений (из 17146)	Accuracy	F-мера
Naive Bayes	5217	0.304	0.309
Random Forest	9091	0.530	0.595

Как видно из табл. 4, увеличение размера выборки не привело к улучшению качества классификации. Результаты для Naive Bayes практически не изменились, а показатели Random Forest существенно просели, едва преодолевая порог в 50 % правильно определенных типов строк, то есть теперь результаты значимо не отличаются от случайного угадывания.

Наконец, на последнем этапе исследования мы решили проверить справедливость принятого ранее допущения, что при разграничении стиха и прозы соизмеряются именно длины строк в слогах. В определении это положение не эксплицировано, а то, что как раз графическая форма «в столбик» позволяет и наивному, и подготовленному читателю правильно распознавать стихотворную форму, является общим местом

стиховедения. К имеющимся признакам, выражающим разницу в длинах строк в слогах мы добавили разницу в длинах строк в буквах. Если качество классификации благодаря добавлению этих признаков вырастет, значит, понятие соизмеримости не следует ограничивать длиной строки в слогах. Таблица 5 демонстрирует получившиеся результаты (ср. таблицу 3). Из нее следует, что добавление признаков, включающих разницу в длинах строк по числу букв, не позволили улучшить работу алгоритмов. Naive Bayes и Support Vector Classification показали еще худшие результаты, а Random Forest остался в пределах тех же значений (классификация сработала лучше для 15 строк, то есть на 0.08 %). Таким образом, принятое в начале исследования допущение, что соизмеримость стиховых сегментов лежит в плоскости числа слогов, по-видимому, верно.

Таблица 5

Результат проверки моделей, основанных на разнице в длинах строк и в слогах, и в буквах

Алгоритм	Число правильных решений (из 17146)	Accuracy	F-мера
Naive Bayes	6380	0.372	0.433
Support Vector Classification	9601	0.559	0.630
Random Forest	15921	0.928	0.954

Кажется, что получившиеся цифры должны быть удовлетворительными для исследователя и не могут устраивать инженера. Применение машинного обучения показало, что соизмеримость стихотворных строк действительно имеет место в реальных текстах, определяется статистически и может быть положена в основу автоматической системы, находящей стихи в текстовом потоке. При использовании алгоритма Random Forest машина способна отыскать закономерности в распределении длин строк в слогах и на основе этих закономерностей правильно классифицировать около 93 % новых строк. В то же время создание настоящей автоматической системы такого назначения не может ограничиться достигнутой точностью работы, и, очевидно, должно включать и иные текстовые признаки, которые позволили бы добиться лучших результатов.

Литература

Гаспаров М.Л. Русский стих начала XX века в комментариях. М.: КДУ, 2004. 312 с.

Плунгян В.А. Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира М.: Изд-во РГГУ, 2011. 672 с.

Шапир М.И. Universum versus: Язык – стих – смысл в рус. поэзии XVIII–XX веков. М.: Языки русской культуры, 2000. Кн. 1. VIII, 536 с.

B. V. Orekhov

National Research University Higher School of Economics
(Russia, Moscow)
nevmenandr@gmail.com

COMMENSURABILITY OF VERSE SEGMENTS: MACHINE LEARNING TEST

This article dicusses the classical problem of the relation between verse and prose in a practical way with the use of machine learning algorithms. We have trained the model on the training sample size of 2300 texts (117 657 prosaic lines and 62 196 verse lines) using such algorithms as Naive Bayes, Random Forest and Support Vector Classification. In one case features were the length of the line and it's neighbors in window 4. In the other case features were the difference of the length between the lines. The test sample consisted of 100 texts and 17146 lines. It turned out that the worst classifier was built on Naive Bayes. Random Forest is able to classify correctly 93 % of the lines. Support Vector Classification showed mean values of accuracy. Another model was built on the base of RNC text collection. It showed the lowest results. That means that the commensurability of lengths of the syllables is indeed an important criterion for the verse lines recognition and definition of the verse itself. The length in the letters does not play a significant role.

Key words: verse, prose, machine learning, naive Bayes, poetic corpus.

References

Gasparov M.L. *Russkii stikh nachala XX veka v kommentariyakh*. [Russian verse of the early XX century in commentaries]. Moscow, KDU Publ., 2004. 312 p.

Plungian V.A. *Vvedenie v grammaticheskuyu semantiku: grammaticheskie znacheniya i grammaticheskie sistemy yazykov mira* [Introduction to grammatical semantics: grammatical meanings and grammatical systems of the world's languages]. Moscow, Russian State Univ. for the Humanities Publ, 2011. 672 p.

Shapir M.I. *Universum versus: Yazyk – stikh – smysl v rus. poezii XVIII–XX vekov* [Universum versus: language – verse – sence in Russian poetry of XVIII–XX centuries]. Moscow, Yazyki Russkoi Kul'tury Publ., 2000. Book 1. VIII, 536 p.