ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 84 за 2024 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Э.С. Клышинский, Н.А. Кочеткова, О.В. Карпик

О некоторых трудностях при определении авторства текста

Рекомендуемая форма библиографической ссылки: Клышинский Э.С., Кочеткова Н.А., Карпик О.В. О некоторых трудностях при определении авторства текста // Препринты ИПМ им. М.В.Келдыша. 2024. № 84. 15 с. https://doi.org/10.20948/prepr-2024-84
https://library.keldysh.ru/preprint.asp?id=2024-84

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Э.С. Клышинский, Н.А. Кочеткова, О.В. Карпик

О некоторых трудностях при определении авторства текста

Москва — 2024

Клышинский Э.С., Кочеткова Н.А., Карпик О.В.

Рассмотрен вопрос изменения взаимного расположения текстов, написанных различными авторами, при уменьшении размеров фрагментов этих произведений. Для векторизации текста использовалась статистика встречаемости синтаксических связей, задаваемых как части речи слов и тип связи. Показано, что уменьшение размеров фрагментов приводит к размытию границ, в рамках которых находятся произведения одного автора.

Ключевые слова: атрибуция авторства, синтаксические характеристики, корпус текстов, русский язык, автоматическая обработка текстов

Klyshinsky E.S., Kochetkova N.A., Karpik O.V.

The paper investigates the issue of changing the relative position of texts written by different authors, while reducing the size of fragments of these works. To vectorize the text, we used statistics on the occurrence of syntactic connections, defined as parts of speech of words and the relation's type. It is shown that decrease of the size of fragments leads to a blurring of the boundaries within which the works of one author are located.

Keywords: attribution of authorship, syntactic characteristics, corpus of texts, the Russian language, automatic text processing

Оглавление

1. Введение

Определение авторства статистическими методами основано на предположении, что, измеряя некоторые текстовые особенности, мы можем различать тексты, написанные разными авторами.

Первые попытки дать количественную оценку стилю письма восходят к XIX веку, когда Менденхолл (1887) исследовал пьесы Шекспира. В первой половине XX века значительный вклад в понимание этой задачи внесли статистические исследования Йоля (1938; 1944) и Ципфа (1932). Позже подробное исследование Мостеллера и Уоллеса (1964) об авторстве «Документов федералиста» (серия из 146 политических эссе, написанных Джоном Джеем, Александром Гамильтоном и Джеймсом Мэдисоном, на двенадцать из которых заявляли права и Гамильтон, и Мэдисон) была, несомненно, самой влиятельной работой в определении авторства, положившей начало «неэкспертным» методам в области определения авторства. Их метод был основан на байесовском статистическом анализе частот небольшого набора слов, относящихся к служебным частям речи (например, «и», «к» и т. д.), и позволил получить значительное улучшение меры различия между кандидатами.

По оценкам Рудмана (1998) было предложено около 1000 различных признаков: длина предложения, длина слова, частота слов, частота символов, размер словарного запаса автора и др. Наиболее характерным примером является метод CUSUM (или QSUM) (Morton & Michealson, 1990), который применялся в суде в качестве экспертного доказательства. Тем не менее, исследовательское сообщество подвергло его серьезной критике и сочло в целом ненадежным (Holmes & Tweedie, 1995). Главной проблемой этого раннего периода было отсутствие объективной оценки, так как анализировались литературные произведения неизвестного или оспариваемого авторства. Также данные были слишком длинными (обычно целые книги) и, вероятно, не были стилистически однородными, число авторов-кандидатов было слишком мало (обычно 2 или 3), данные были не однородны по тематике. Оценка предложенных методов обычно была основана на субъективном визуальном осмотре диаграмм рассеяния. Сравнение различных методов было невозможно из-за отсутствия эталонных данных.

Помимо задачи определения авторства существуют и смежные, к которым относятся проверка авторства (нужно определить, был ли данный текст написан определенным автором) (Koppel & Schler, 2004), нахождение плагиата (то есть нахождение сходства между двумя текстами) (Meyer zu Eissen, Stein & Kulig, 2007; Stein & Meyer zu Eissen, 2007), профилирование автора (извлечение информации о возрасте, образовании, поле и других характеристиках автора данного текста) (Koppel, Argamon, & Shimoni, 2002), обнаружение стилистических несоответствий (например, в работах, написанных в соавторстве) (Collins, Kaufer, Vlachos, Butler, & Ishizaki, 2004).

В классической задаче определения авторства тексту неизвестного автора должен быть поставлен в соответствие автор из известного множества авторов и сопоставленных им текстов. С точки зрения машинного обучения здесь применима многоклассовая задача классификации текстов (Sebastiani, 2002). Соответственно, её сложность нелинейно растёт при увеличении числа авторов и их текстов. С точки зрения статистики, достоверность результатов падает при уменьшении размера текста.

В данной работе мы рассмотрим изменение поведения системы при небольшом увеличении числа авторов. Главной задачей работы является исследование степени сходства в рамках произведений одного автора, когда размер текста уменьшается от романа или повести до нескольких страниц.

2. Обзор существующих решений

Первые попытки определить авторство были основаны на простых мерах, таких как количество предложений и количество слов, которые могут применяться к любому языку и любому корпусу без дополнительных требований кроме наличия токенизатора.

Характеристика емкости словарного запаса – это попытка количественно оценить разнообразие использованных слов в тексте. Типичными примерами являются тип-токен отношение (отношение размера словаря текста к его длине в токенах) и количество hapax legomena (т.е. слов, встречающихся один раз) (de Vel, Anderson, Corney, & Mohay, 2001). Так как размер словаря нелинейно зависит от длины текста (словарь увеличивается быстро вначале, а затем скорость падает), были предложены методы для достижения независимости от длины текста (Yule, 1944; Honore, 1979).

Самый распространенный подход к представлению текстов – это вектор частот слов или «мешок слов». Текст рассматривается как набор слов, каждое из которых имеет частоту встречаемости без какого-либо учета контекстной информации. Более частотные слова (артикли, предлоги, местоимения и др. функциональные части речи) являются одной из лучших характеристик в задаче определения авторов (Burrows, 1987; Argamon & Levitan, 2005). В таких задачах, как тематическая классификация текста, подобные слова обычно, напротив, исключаются из рассмотрения, поскольку они не несут семантической информации. Функциональные слова используются авторами в значительной степени бессознательно и не зависят от темы. Таким образом, они могут выявить стиль автора независимо от темы рассматриваемых текстов. Выбор конкретных функциональных слов, которые будут использоваться в качестве признаков, обычно основан на произвольных критериях и требует знания языка.

Простой и довольно успешный метод определения набора лексических характеристик для определения авторства состоит в том, чтобы извлечь наиболее часто встречающиеся слова из корпуса (по всем текстам авторов-кандидатов). Затем необходимо принять решение о количестве или пороговой

частоте, для определения числа часто используемых слов, которые будут использоваться в качестве характеристик. В более ранних исследованиях это количество не превышало 100 часто встречающихся слов (Burrows, 1992).

Парадигма «мешок слов» дает простое и эффективное решение, но не учитывает порядок слов (т.е. контекстную информацию). Чтобы воспользоваться контекстной информацией, в качестве текстовых элементов были предложены n-граммы слов (n идущих подряд слов). Однако точность классификации, основанной на n-граммах, не всегда лучше, чем на отдельных словах (Sanderson & Guenther, 2006; Coyotl-Morales, et al., 2006).

Koppel и Schler (2003) предложили учитывать ошибки в написании слов как особенности авторского стиля. Для этого они определили набор орфографических ошибок (например, пропуски букв и вставки) и ошибок форматирования (например, все прописные слова) и предложили метод автоматического извлечения слов с ошибками с использованием средств проверки орфографии.

Более сложным способом представления текста является использование синтаксической информации. Авторы неосознанно склонны использовать одни и те же синтаксические паттерны. Поэтому синтаксическая информация считается более надежным авторским маркером по сравнению с лексической. Кроме того, функциональные слова в представлении стиля также обычно встречаются в определенных синтаксических конструкциях. Baayen, van Halteren и Tweedie (1996) первыми использовали синтаксическую информацию для определения авторства. Основываясь на синтаксически аннотированном английском корпусе, включающем полуавтоматически созданное полное дерево разбора каждого предложения, они смогли извлечь частоты правил перехода. Подобная информация описывает как синтаксический класс каждого слова, так и то, как слова объединяются во фразы или иные структуры. Эксперименты показали, что этот тип характеристик дает более точные результаты, чем измерение словарного запаса и лексические показатели.

Еще одна попытка использовать синтаксическую информацию была предложена Stamatatos, et al. (2000; 2001). Использовалась информация о границах предложений и синтаксических групп, эксперименты проводились на текстах на современном греческом языке. Извлекались такие характеристики, как количество групп существительного, количество групп глагола, длина групп существительного, длине групп глагола и т.д.

Sidorov, et al. (2014) использовали синтаксические n-граммы (sn-граммы), где элементы выбираются не по их порядку появления в тексте, а по их положению в синтетическом дереве. Эксперимент показал, что sn-граммы делают акцент на синтаксических отношениях между словами. Были использованы sn-граммы слов нескольких типов меток: часть речи, слова, смешанные sn-граммы.

Для извлечения ряда характеристик достаточно частичного синтаксического анализа (Luyckx & Daelemans, 2005; Uzuner & Katz, 2005).

Graham, N., Hirst, G., & Marthi (2005) преобразовали вывод частичного парсера в упорядоченный поток синтаксических меток. Затем они подсчитали частоты биграмм меток из этого потока, чтобы представить контекстную синтаксическую информацию, и обнаружили, что эта информация полезна для различения авторов очень коротких текстов (длиной около 200 слов).

Еще более простой подход состоит в том, чтобы использовать только метки части речи. Для этого необходим морфологический анализатор, инструмент, который присваивает метки морфосинтаксической информации каждому токену на основе контекстной информации.

Koppel and Schler (2003) предложили также использовать синтаксические ошибки. Для обнаружения такой информации они использовали программу проверки орфографии. К сожалению, средства проверки орфографии не очень точны, и авторам пришлось изменить выходные данные этого инструмента, чтобы улучшить результаты обнаружения ошибок.

Koppel, et al. (2011) предложили использовать синтаксические шаблоны, сформированные из частотных поддеревьев дерева разбора. Они показали, что информация, содержащаяся в различных типах синтаксических признаков, лучше всего сочетается при использовании правила Демпстера по сравнению с некоторыми другими методами слияния информации. Предложенный алгоритм показал свою надежность при классификации постов из блогов 100 авторов.

Заметим, что все эти авторы использовали текст как единицу анализа. Хотя размер текста по экспериментам мог существенно отличаться, в них не исследовалось, насколько один автор внутри разных своих текстов может отличаться по набору параметров. Как уже отмечалось выше, для идентификации автора необходимо тщательно подбирать тексты по стилям, при этом внутри одного произведения автор может чередовать литературные приёмы. Например, «Чёрный отряд» Глена Кука написан от лица разных героев с полной сменой стилистики изложения.

В связи с этим встаёт проблема исследования постоянства значений из набора авторских характеристик. В данной работе мы будем исследовать постоянство статистики синтаксических связей как метода, показавшего лучшие результаты.

3. Метод исследования и материалы

В качестве пространства признаков мы выбрали синтаксические связи между словами. Связь характеризуется как тройка: часть речи главного слова, часть речи зависимого слова и тип связи. Все произведения должны быть проанализированы с использованием одной библиотеки синтаксического анализа, множество троек должно быть объединено для формирования единого пространства признаков. Далее рассчитываются частоты встречаемости связей.

Для исследования постоянства значений характеристик авторского стиля разобьем произведение на фрагменты по 10-100 предложений. Характеристики

рассчитываются для каждого фрагмента по отдельности. Векторизация производится в пространстве объединённых характеристик.

Отдельно будут рассматриваться такие характеристики, как доля глаголов, прилагательных и существительных во фрагменте или произведении.

Исследование проводилось на двух коллекциях. В состав первой коллекции входят победители премии «Большая книга» (Н. Абгарян, С. Беляков, Е. Веркин, Ш. Идиатуллин, А. Ильичевский, А. Макушинский, А. Поляринов, А. Сальников, М. Степнова, Г. Яхина) и фрагменты корпуса СинТагРус. В состав второй коллекции входят те же победители премии, но с дополнительными произведениями, не получившими премию, и произведения научной фантастики (Беляев, Бушков, Мошков, Преображенский, Регентов, Шушпанов). Часть произведений представлена ознакомительными фрагментами.

4. Результаты экспериментов

Для первой коллекции проведём следующий эксперимент. Разобьем произведения на фрагменты по 100 предложений. Для каждого фрагмента рассчитаем относительную долю глаголов, прилагательных и существительных. Для каждого из произведений у нас получится три распределения значений, по одному распределению для из параметров. Рассчитаем доверительный интервал для этих трёх параметров для каждого из произведений.

Далее построим график, на котором по оси абсцисс будет откладываться доля существительных во фрагменте, а по оси ординат – доля глаголов или прилагательных. Помимо этого, отметим значения, посчитанные для произведения в целом и их доверительные интервалы. Результаты показаны на рис. 1 и 2.

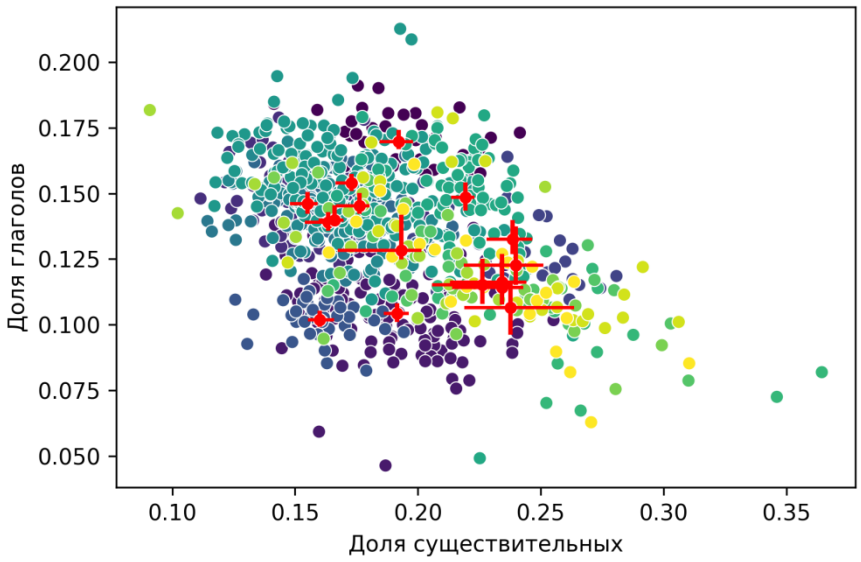

Рис. 1. Распределение доли глаголов и существительных по фрагментам и произведениям. Значения для произведений и их доверительные интервалы отмечены красным

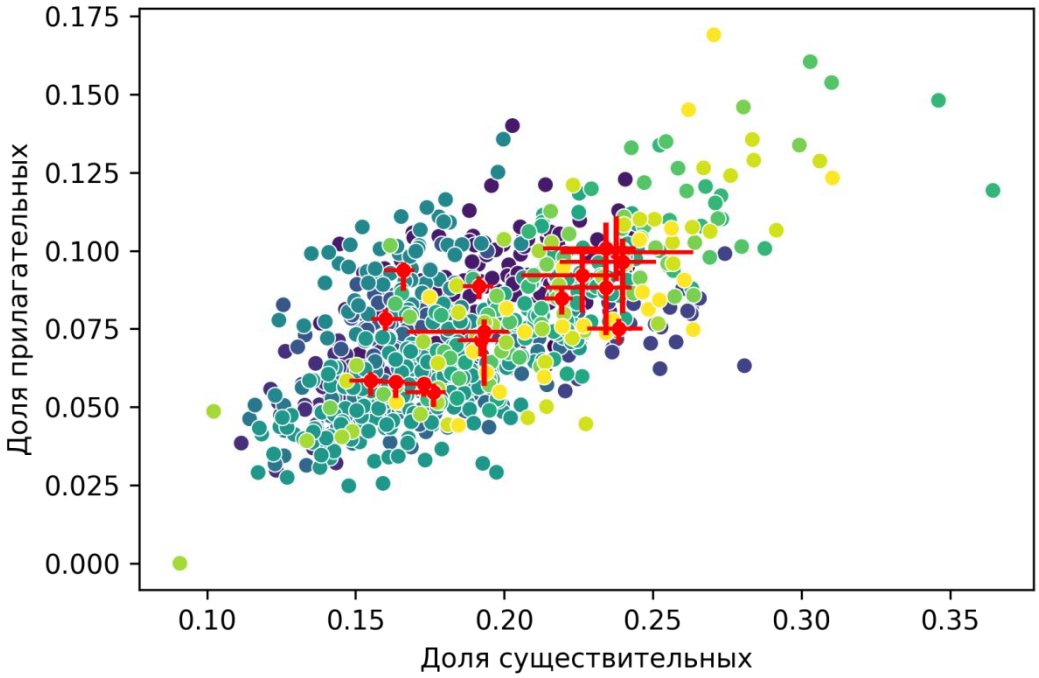

Рис. 2. Распределение доли прилагательных и существительных по фрагментам и произведениям. Значения для произведений и их доверительные интервалы отмечены красным

Из рисунков видно, что использование только этих параметров не позволяет разделить авторов произведений. Более того, уменьшение размеров фрагментов приводит к полному смешению и невозможности определения авторов.

Особое внимание стоит уделить расположению точек, отмечающих произведения в целом и их доверительных интервалов (на рисунке отмечены красным). Для большей части произведений доверительный интервал оказывается сдвинут относительно произведения в целом. Это означает, что среднее по фрагментам не складывается в среднее по произведению. Следовательно, во фрагментах произведения наблюдается существенный сдвиг значений параметров. Например, это может быть объяснено наличием нескольких фрагментов, существенно отличающихся в одну сторону, тогда как произведение в целом написано относительно ровным стилем.

Наше предположение подтверждается изучением материала. Произведение, для которого наблюдается самый большой сдвиг, было взято из корпуса СинТагРус и содержит в себе фрагменты, написанные разными авторами. Из-за этого наблюдается существенное изменение значений параметров между фрагментами. Впрочем, несимметричность расположения доверительного интервала присуща и файлам, содержащим в себе тексты одного автора.

Итак, мы можем утверждать, что как минимум некоторые из параметров произведения претерпевают существенное изменение по ходу развития повествования. Это наблюдение соответствует известным данным. Так, например, известно, что доля глаголов в произведении отражает темп повествования. Она увеличивается в главах, содержащих более активные действия, и уменьшается в главах, содержащих описания.

Теперь перейдём к анализу второй коллекции текстов, содержащей в себе несколько произведений для некоторых авторов. Рассчитаем статистику синтаксических связей для каждого из произведений в целом, векторизуем её и отобразим на плоскость при помощи метода UMAP. Результат показан на рис.3.

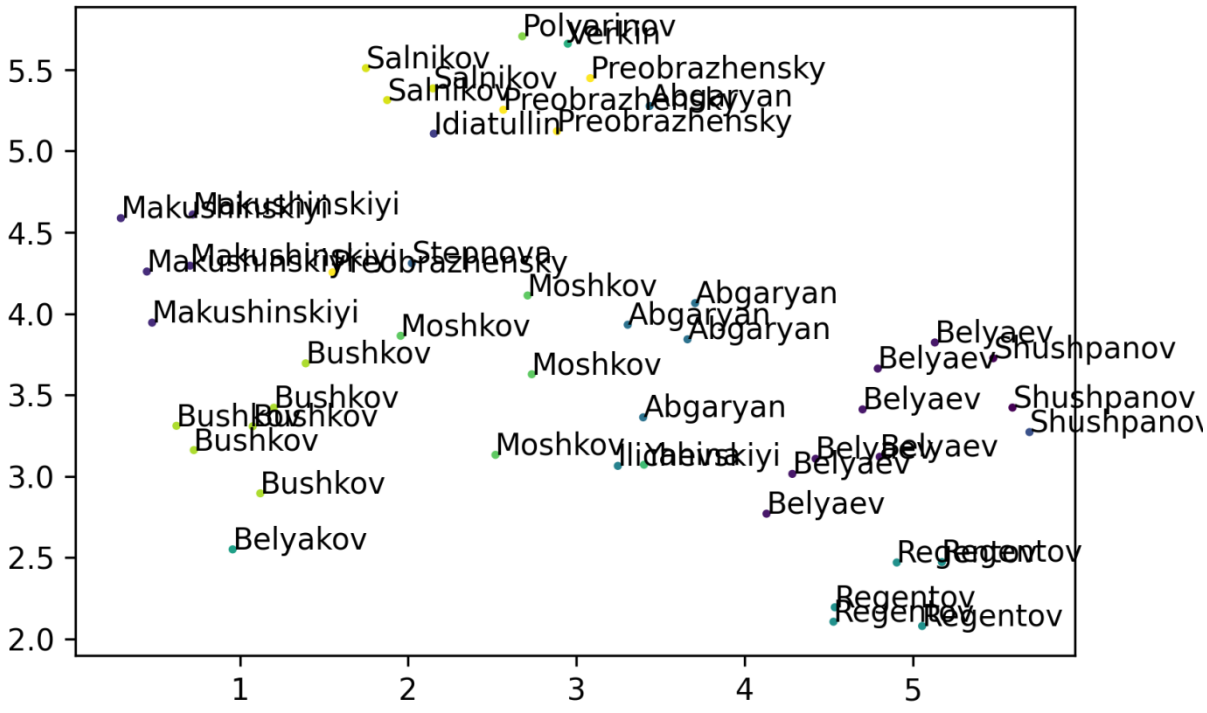

Рис. 3. Группировка произведений по статистике синтаксических связей

На рис. 3 видно, что произведения хорошо группируются по авторам. Единственным исключением является одно из произведений Преображенского, тяготеющее к произведениям Макушинского, Бушкова и Мошкова. Более того, одно из произведений Шушпанова написано в соавторстве с Лукьяненко по вселенной Дозоров, однако наши результаты показывают, что соавторство является скорее номинальным и отражает права на использование героев.

Теперь проведём аналогичную процедуру, разделив произведения на фрагменты по 100 предложений. Заметим, что это относительно большой объем, соответствующий примерно 7-10 страницам текста, то есть короткому рассказу. Результаты разделения показаны на рис. 4. Здесь видно, что разделимость фрагментов и произведений сохраняется, то есть метод может служить для определения авторства. Однако детальный анализ демонстрирует смешивание некоторых авторов. Так, одно из произведений Преображенского смещается к произведениям Бушкова, а одно из произведений Макушинского отделяется от основной группы (в данном случае возможны артефакты UMAP, однако сходный результат был получен при разных прогонах).

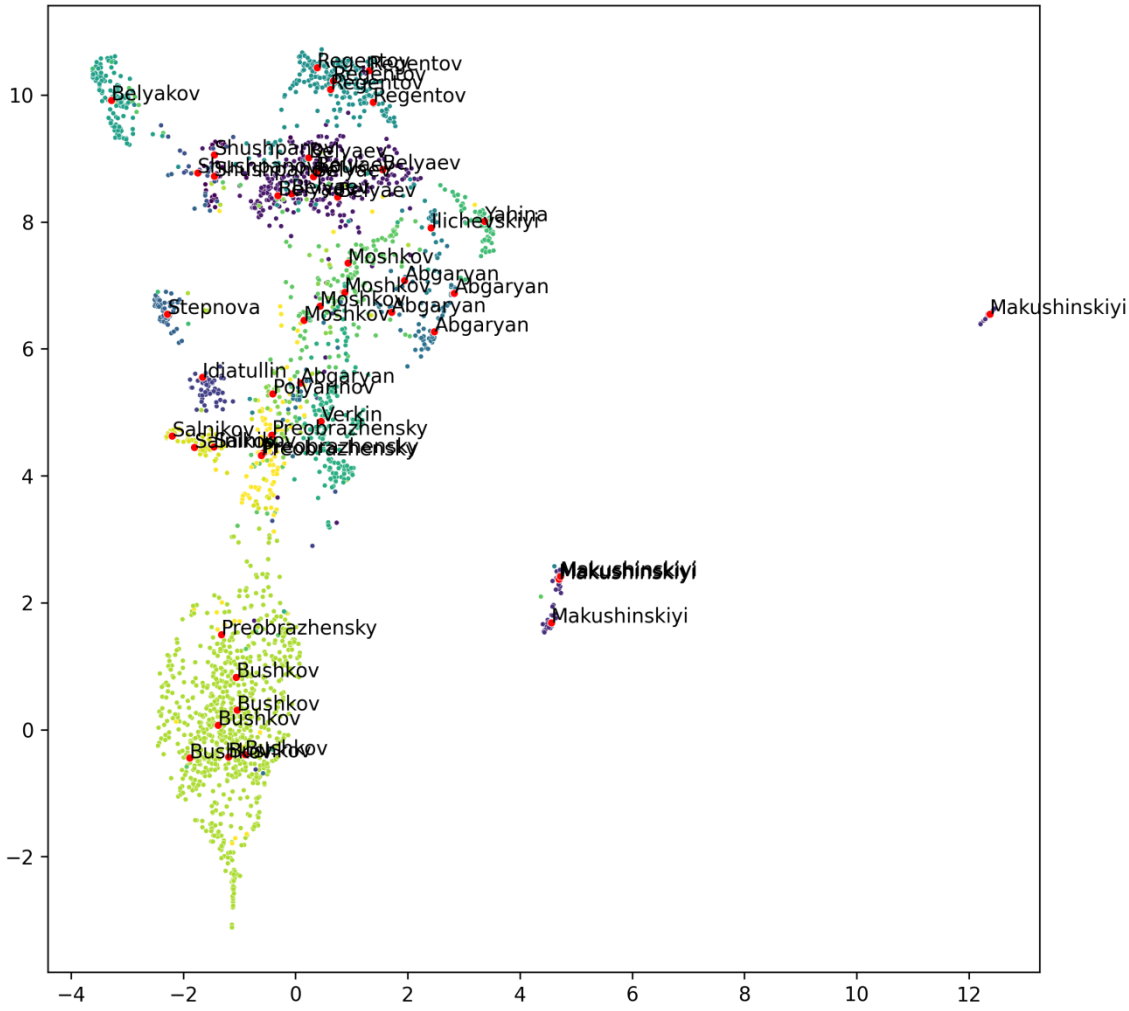

Рис. 4. Группировка фрагментов произведений по статистике синтаксических связей, фрагменты по 100 предложений

Далее мы уменьшили размер фрагмента до 50 предложений (около 4 страниц текста). Результаты показаны на рис. 5. На рисунке видно, что смешение между некоторыми авторами увеличилось. Так, Абгарян, Мошков и Беляев смешиваются на рисунке в нескольких областях. При снижении размера фрагмента до 20 предложений (1-2 страницы) авторы полностью перемешиваются. Малое количество примесей наблюдается лишь в зоне, занятой произведениями Бушкова, и для одного из произведений Макушинского.

Полученные эффекты могут также быть артефактами, вызванными снижением размерности пространства. В связи с этим мы рассчитали косинусное расстояние между произведениями некоторых авторов, приведённое в Табл. 1. Из таблицы видно, что среднее расстояние между фрагментами могло бы служить в качестве некоторого порогового значения, позволяющего разделять фрагменты авторов (доля пар фрагментов, находящихся дальше среднего расстояния, не превышает 40%). При этом для

рассмотренных пар авторов всегда существует пара написанных ими фрагментов настолько близких, что авторство не может быть определено только исходя из использованных синтаксических параметров. С другой стороны, косинусная близость между средними значениями, рассчитанными по произведению целиком, не превышает 0.1531, а среднее расстояние равно 0.042. При этом возникает гораздо меньше смешиваний произведений целиком, а применение метода UMAP позволяет улучшить классификацию.

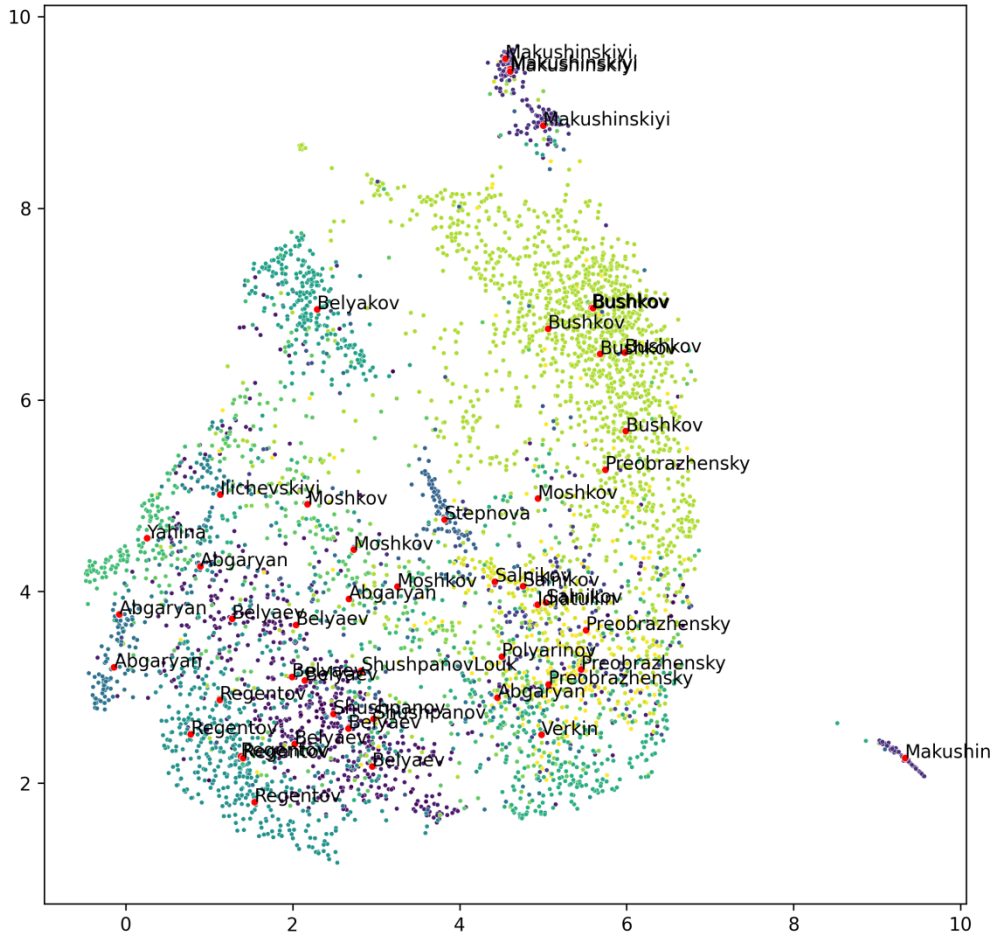

Рис. 5. Группировка фрагментов произведений по статистике синтаксических связей, фрагменты по 50 предложений

Для сравнения значений мы рассчитали среднее, минимальное и максимальное расстояние между точками внутри произведений одного автора (см. Табл. 2). Из таблицы видно, что среднее расстояние между фрагментами, написанными одним автором, может быть сопоставимо со средним расстоянием между фрагментами, написанными разными авторами (ср. данные для Бушкова, для которого включено несколько произведений). Максимальное расстояние между фрагментами также может быть сопоставимым (ср. данные для Белякова, для которого также включено несколько произведений).

Таблица 1.

Косинусные расстояния между фрагментами авторов

Автор 1	Автор 2	Ср. расст.	Min расст.	Max расст.
Бушков	Абгарян	0.156	0.028	0.798
Бушков	Мошков	0.145	0.035	0.865
Бушков	Макушинский	0.152	0.034	0.7197
Беляков	Макушинский	0.154	0.036	0.885
Беляков	Бушков	0.1685	0.043	0.924
Беляков	Преображенский	0.1879	0.041	0.914
Абгарян	Яхина	0.0959	0.017	0.728
Абгарян	Макушинский	0.137	0.037	0.757
Абгарян	Сальников	0.095	0.023	0.764

Таблица 2.

Косинусные расстояния между фрагментами одного автора

Автор 1	Ср. расст.	Min расст.	Max расст.
Абгарян	0.073	0.013	0.723
Бушков	0.154	0.011	0.748
Беляков	0.103	0.024	0.856
Макушинский	0.084	0.009	0.403
Мошков	0.106	0.022	0.633
Преображенский	0.096	0.017	0.491
Сальников	0.072	0.015	0.332
Яхина	0.074	0.015	0.466

5. Заключение

Анализ полученных данных позволяет говорить о том, что различия в стиле изложения одного автора, измеренном при помощи статистики употребления синтаксических связей, может оказаться сопоставим с различиями между авторами, если измерения проводятся не по произведению объёмом несколько сотен страниц, а по фрагментам в 5-10 страниц. Чем меньше объем фрагментов, тем больше будет смешивание авторских стилей, и, следовательно, меньше шансов на корректную классификацию авторов.

Заметим, что здесь также необходимо обращать внимание на стиль, которым пишется произведение в целом. Так, диалоги могут описываться как с авторскими словами, так и без них. Диалоги без авторских слов часто пишутся короткими рублеными фразами, что существенно изменяет статистику употребления синтаксических связей или долю слов определенных частей речи (например, в подобных диалогах относительно редки прилагательные). Можно предположить, что фрагменты, содержащие в себе подобные диалоги, будут находиться близко друг от друга. С другой стороны, авторские предпочтения, выражающиеся в применении или отсутствии междометий, союзов, служебных

глаголов, могут увеличить расстояние между фрагментами, что также найдёт своё отражение в рассчитываемой статистике.

Можно предположить, что вместо того, чтобы рассматривать произведение как единую точку, следует перейти к анализу распределений, получаемых на фрагментах одного произведения. Помимо этого возможен расчёт информации для отдельных фрагментов произведения с известным авторством и дальнейшая атрибуция фрагментов нового произведения. Подобный подход позволит оперировать вычислением вероятности сходства большинства фрагментов взамен поиска единственного наиболее сходного произведения.

Список литературы

1. Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing.
2. Baayen, R., van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing, 11(3), 121-131.
3. Burrows, JF (1987). Word patterns and story shapes: The statistical analysis of narrative style. Literary and Linguistic Computing, 2, 61-70.
4. Burrows, JF (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. Literary and Linguistic Computing, 7(2), 91-109.
5. Collins, J., Kaufer, D., Vlachos, P., Butler, B., & Ishizaki, S. (2004). Detecting collaborations in text: Comparing the authors' rhetorical language choices in the Federalist Papers. Computers and the Humanities, 38, 15-36.
6. Coyotl-Morales, RM, Villaseñor-Pineda, L., Montes-y-Gómez, M., & Rosso, P. (2006). Authorship attribution using word sequences. In Proceedings of the 11th Iberoamerican Congress on Pattern Recognition (pp. 844-853) Springer.
7. Graham, N., Hirst, G., & Marthi, B. (2005). Segmenting documents by stylistic character. Journal of Natural Language Engineering, 11(4), 397-415.
8. Holmes, DI, & Tweedie, FJ (1995). Forensic stylometry: A review of the cusum controversy. In Revue Informatique et Statistique dans les Sciences Humaines. University of Liege (pp. 19-47).
9. Koppel, M., Argamon, S., & Shimoni, AR (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), pp. 401-412.
10. Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis (pp. 69-72).
11. Mendenhall, TC (1887). The characteristic curves of composition. Science, IX, 237-49.

12. Meyer zu Eissen, S., Stein, B., & Kulig, M. (2007). Plagiarism detection without reference collections. Advances in Data Analysis (pp. 359-366) Springer.
13. Mosteller, F. & Wallace, DL (1964). Inference and disputed authorship: The Federalist. Addison-Wesley.
14. Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities, 31, 351-365.
15. Sanderson, C., & Guenter, S. (2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In Proceedings of the International Conference on Empirical Methods in Natural Language Engineering (pp. 482-491).
16. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1).
17. Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. Computational Linguistics, 26(4), 471-495, 2000.
18. Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. Computers and the Humanities, 35(2), 193-214.
19. Stein, B., & Meyer zu Eissen, S. (2007). Intrinsic plagiarism analysis with meta learning. In Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection (pp. 45-50).
20. Uzuner, O., & Katz, B. (2005). A comparative study of language models for book and author recognition. In Proceedings of the 2nd International Joint Conference on Natural Language Processing (pp. 969-980) Springer.
21. de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. SIGMOD Record, 30(4), 55-64.
22. Yule, GU (1944). The statistical study of literary vocabulary. University Press, Cambridge.
23. Zipf, GK (1932). Selected studies of the principle of relative frequency in language. Harvard University Press, Cambridge.
24. Sidorov, G.; Velasquez, F.; Stamatatos, E.; Gelbukh, A.; and Chanona-Hernandez, L (2014) Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications 41(3):853–860.
25. Luyckx, K., Daelemans, W.: Authorship Attribution and Verification with Many Authors and Limited Data. In: Proc. of the Twenty-Second International Conference on Computational Linguistics (COLING 2008), Manchester, UK (2008) 513-520.
26. Koppel, Moshe & Schler, Jonathan & Argamon, Shlomo. (2011). Authorship attribution in the wild. Language Resources and Evaluation. 45. 83-94. 10.1007/s10579-009-9111-2.