Формальные методы определения авторства текстов

Article · January 2012

CITATION
1

READS
540

1 author:

Tatiana Batura
A.P. Ershov Institute of Informatics Systems
35 PUBLICATIONS **57** CITATIONS

SEE PROFILE

УДК 519.68; 681.513.7; 612.8.001.57; 007.51.52

Т. В. Батура

Институт систем информатики им. А. П. Ершова СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

E-mail: tbatura@ngs.ru

ФОРМАЛЬНЫЕ МЕТОДЫ ОПРЕДЕЛЕНИЯ АВТОРСТВА ТЕКСТОВ

Представлен обзор формальных методов установления авторства (атрибуции) текстов. В статье приведено описание наиболее известных программных систем для определения авторского стиля, ориентированных на русский язык, предпринята попытка произвести их сравнительный анализ, выявить особенности и недостатки рассмотренных подходов. При решении задачи определения авторства текстов наибольший интерес и наибольшую сложность представляет анализ синтаксического, лексико-фразеологического и стилистического уровней текста. Экспертный анализ авторского стиля является трудоемким процессом, поэтому в работе уделяется внимание именно формальным методам идентификации автора текста. В настоящее время для атрибуции текстов применяются подходы из теории распознавания образов, математической статистики и теории вероятностей, алгоритмы нейронных сетей, кластерного анализа и др. Среди проблем, затрудняющих исследования в области атрибуции, можно выделить проблему выбора лингвостилистических параметров текста и составления выборки эталонных текстов. Необходимо проводить дальнейшие исследования, направленные на поиск новых или совершенствование уже имеющихся методов атрибуции текстов, поиск характеристик, позволяющих четко разделять стили авторов, в том числе на коротких текстах и на малых объемах выборки.

Ключевые слова: атрибуция текста, определение авторства, формальные параметры текста, авторский стиль, классификация текстов.

Введение

Для определения автора текста зачастую приходится обращаться к экспертам. Эксперты могут идентифицировать автора неизвестного текста или определить принадлежность произведения другому автору при помощи характерных языковых особенностей, стилистических приемов. Несомненно, экспертный анализ авторского стиля является трудоемким процессом, поэтому в данной статье уделяется внимание именно формальным методам определения автора. Важно отметить, что задача установления авторства текстов (задача атрибуции) встречается в различных областях и представляет интерес для филологов, литературоведов, юристов, криминалистов, историков. Поэтому появляется потребность в создании формальных методов ее решения. В настоящее время для атрибуции текстов применяются подходы из теории распознавания образов, математической статистики и теории вероятностей, алгоритмы нейронных сетей и кластерного анализа и многие другие.

С развитием вычислительной техники появилась возможность реализовать методы, требующие огромных вычислений, чтобы облегчить работу эксперов. Существующие программные продукты позволяют учитывать и варьировать различные лингвостатистические параметры, характеризующие текст с разных сторон. В статье приведен обзор различных формальных методов определения авторства текстов, предпринята попытка выявить особенности и недостатки рассмотренных методов, сравнить программные продукты по атрибуции текстов на русском языке.

Лингвостатистические параметры анализируемого текста

Атрибуция текста – исследование текста с целью установления авторства или получения каких-либо сведений об авторе и условиях создания текстового документа. Задачи атрибуции можно разделить на идентификационные и диагностические [1].

Идентификационные задачи позволяют осуществить проверку авторства:

- подтвердить авторство определенного лица;
- исключить авторство определенного лица;
- проверить тот факт, что автором всего текста был один и тот же человек;
- проверить тот факт, что написавший текст является при этом его настоящим автором.

Идентификационные задачи решаются из предположения, что автор текста известен.

Диагностические задачи позволяют определить личностные характеристики автора (образовательный уровень, родной язык, знание иностранных языков, происхождение, место постоянного проживания и др.) и / или факт сознательного искажения письменной речи. Диагностические задачи решаются из предположения, что автор текста неизвестен. В этих случаях обычно невозможно сопоставить исследуемый текст с текстами автора.

Методы атрибуции позволяют исследовать текст на пяти уровнях: пунктуационном, орфографическом, синтаксическом, лексико-фразеологическом, стилистическом.

Пунктуационный уровень помогает выявить особенности употребления автором знаков препинания, характерные ошибки.

Орфографический уровень выявляет характерные ошибки в написании слов.

Синтаксический уровень позволяет определить особенности построения предложений, предпочтение тех или иных языковых конструкций, употребление времен, активного или пассивного залога, порядок слов, характерные синтаксические ошибки.

Лексико-фразеологический уровень определяет словарный запас автора, особенности использования слов и выражений, склонность к употреблению редких и иностранных слов, диалектизмов, архаизмов, неологизмов, профессионализмов, арготизмов, навыки употребления фразеологизмов, пословиц, поговорок, «крылатых выражений» и т. д.

Стилистический уровень позволяет определить жанр, общую структуру текста, для литературных произведений – сюжет, характерные изобразительные средства (метафора, ирония, аллегория, гипербола, сравнение), стилистические фигуры (градация, антитеза, риторический вопрос и т. д.), другие характерные речевые приемы.

Под «авторским стилем» обычно понимаются последние три уровня. Анализ именно синтаксического, лексико-фразеологического и стилистического уровней представляет наибольший интерес и наибольшую сложность.

Существует довольно много методов анализа стиля. В целом можно разделить их на две большие группы – экспертные и формальные. Экспертные методы предполагают исследование текста профессиональным лингвистом-экспертом. К формальным относятся приемы из теории вероятностей и математической статистики, алгоритмы кластерного анализа и нейронных сетей. Наиболее полная классификация основных формальных методов атрибуции текстов дана, например, в работе [1] (см. рисунок). Как видно, формальные методы чаще всего основаны на сравнении вычислимых характеристик текстов, как в теории распознавания образов. Применение теории распознавания образов в задаче атрибуции текстов можно встретить, например, в [1; 2]. В общем случае текст отображается в вектор вычисленных для него параметров, каждый из которых объективно характеризует некоторый набор особенностей текста. Таким образом, текст графически отображается в некоторую точку n-мерного пространства. При такой формализации автор также может быть представлен в виде аналогичного вектора параметров – этим вектором будет вектор текстов, написанных данным автором.

[1] Автороведческая экспертиза. URL: http://ru.wikipedia.org/wiki/Автороведческая_экспертиза; *Галяшина Е. И.* Лингвистическая безопасность речевой коммуникации // ГЛЭДИС. 2004. URL: http://www.rusexpert.ru/magazine/034.htm

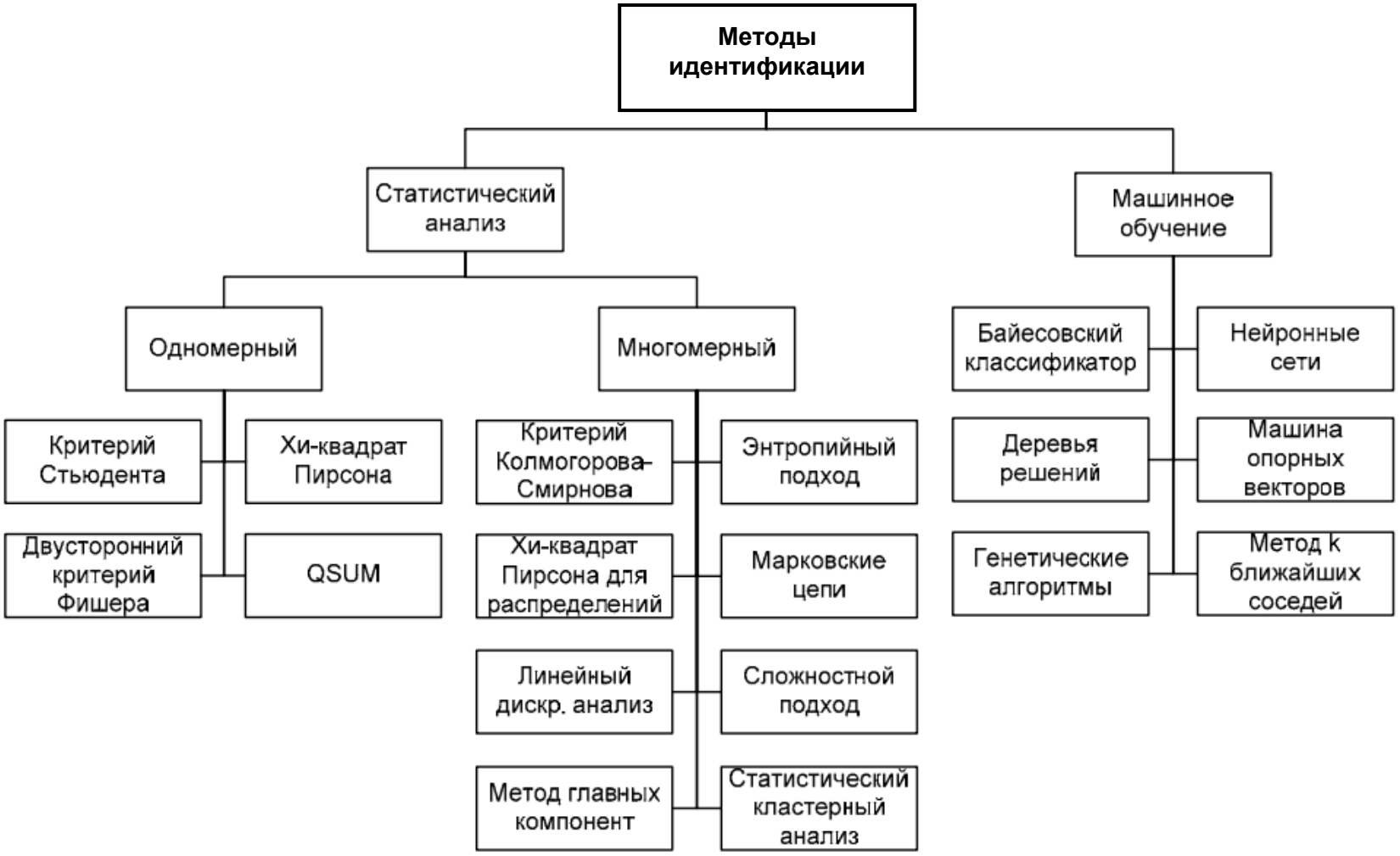

Рис. 1. Формальные методы атрибуции текстов

В качестве критерия близости двух текстов вводится так или иначе вычисляемое «расстояние» между соответствующими векторами. В простейшем случае можно представить наборы параметров как обычные векторы в n-мерном декартовом пространстве, выходящие из начала координат, и считать расстоянием между текстами обычное декартово расстояние между концами соответствующих им векторов. Есть множество других вариантов. Именно «расстояние» является в итоге интегральной характеристикой различия текстов. Оно определенным образом нормируется, и тексты, для которых расстояние велико, считаются с высокой вероятностью относящимися к разным авторам. Таким образом, чтобы сопоставить авторство двух текстов, достаточно вычислить для них параметры и определить расстояние. Чтобы сопоставить текст с автором, сравниваются векторы параметров автора и данного текста, т. е. фактически снова сравниваются два текста – текст с заведомо известным автором (эталонный текст) и текст, авторство которого требуется установить, подтвердить или опровергнуть (спорный текст). Можно также составить векторы формальных параметров, различающие не конкретных авторов (или их группы), а выделяющие определенные характеристики авторов (например, образовательный уровень).

В большинстве случаев в качестве характеризующих параметров текста выбираются те или иные его статистические характеристики: количество использования определенных частей речи, некоторых конкретных слов, знаков препинания, фразеологизмов, архаизмов, редких и иностранных слов, количество и длина предложений (измеренная в словах, слогах, знаках), объем словаря, количество полнозначных и служебных слов, средняя длина предложения, отношение числа глаголов к общему количеству словоупотреблений в тексте и т. д.

Основная проблема формальных методов анализа авторства состоит как раз в выборе параметров. Как было отмечено А. А. Марковым [3], существует целый ряд формальных статистических характеристик текстов, непригодных для определения авторства в силу одного из двух недостатков.

Отсутствие устойчивости. Разброс значений параметра для текстов одного и того же автора настолько велик, что диапазоны возможных значений для разных авторов перекрываются. Очевидно, данный параметр не поможет различать авторов, а при использовании в составе группы параметров лишь сыграет роль дополнительного шума.

Отсутствие различающей способности. Параметр может принимать близкие значения для всех или большинства авторов, поскольку его значение определяются свойствами языка, на котором написаны тексты, а не индивидуальными особенностями создателя текста. Поэтому параметры, используемые в формальных методиках определения авторства, должны предварительно исследоваться на устойчивость и различающую способность, желательно на текстах большого количества различных авторов. В работе [4] выделены следующие три условия применимости формального параметра.

Массовость. Параметр должен опираться на те характеристики текста, которые слабо контролируются автором на сознательном уровне. Это необходимо, чтобы устранить возможность сознательного искажения автором характерного для него стиля или имитации стиля другого автора.

Устойчивость. Параметр должен сохранять постоянное значение для одного автора. Естественно, в силу случайных причин некоторое отклонение значений от среднего неизбежно, но оно должно быть достаточно мало.

Различающая способность. В идеале параметр должен принимать существенно различные значения (превышающие колебания, возможные для одного автора) для разных авторов. Необходимо отметить, что выбрать параметры, которые гарантированно разделяют двух любых авторов, очень трудно. Какими бы ни были параметры, всегда существует вероятность того, что два или более автора окажутся по данным параметрам близки в силу случайного совпадения. Поэтому на практике считается достаточным, чтобы параметр позволял уверенно различать между собой разные группы авторов, т. е. существовало достаточно большое количество групп авторов, для которых средние значения параметра значительно различаются. Параметр, очевидно, не поможет различить тексты авторов из одной группы, но позволит уверенно различать тексты авторов, попавших в разные группы. Различать тексты авторов одной группы можно за счет использования одновременно достаточно большого вектора различных по характеру параметров – в этом случае вероятность случайного совпадения ста-

нет заметно меньше. Для уверенного вывода в отношении текстов, для которых формально вычисленное параметрическое расстояние мало, требуется дополнительное исследование экспертными методами.

Система «Лингвоанализатор»

Ряд исследований был проведен Д. В. Хмелёвым [5; 6], результатом которых явился вывод об эффективности применения алгоритмов сжатия данных для задачи определения авторства. Также был сделан вывод о том, что простейший подход с использованием цепей Маркова первого порядка показывает хорошие результаты на файлах большого объема и плохие по сравнению с другими методами на отрывках длиной в 2 000–5 000 символов. Этот метод был реализован в системе «Лингвоанализатор» (http://www.rusf.ru/books/analysis).

Существенное преимущество метода энтропийной классификации (с помощью сжатия) состоит в отсутствии предварительной обработки текста. Суть метода в том, чтобы добавлять текст, автор которого неизвестен, к тексту, принадлежащему конкретному автору, и смотреть, насколько хорошо сжимается эта «добавка». Правильный исходный класс документа – это тот, на котором он жмется лучше всего.

Метод, предложенный Д. В. Хмелёвым, основан на применении относительной энтропии. Есть несколько способов вычислить эту характеристику: «шельфовый» (off-the-shelf) алгоритм, метод предсказания по частичному совпадению (PPM – Prediction by Partial Matching) и использование индекса повторяемости.

Среди алгоритмов сжатия данных без потерь наиболее часто встречающимися являются: кодирование Хаффмана, арифметическое кодирование, метод Барроуза – Уилера и множество вариаций метода Лемпеля – Зива (LZ). К алгоритмам, специально ориентированным на сжатие текста, относятся следующие: PPM использует Марковскую модель небольшого порядка и DMC (Dynamic Markov compression) использует динамически изменяемую Марковскую модель. В рамках подхода PPM правильный исходный класс документа – это тот, на чьей модели получается наилучшее сжатие. Каждый алгоритм имеет большое число модификаций и параметров (например, существует динамическое кодирование Хаффмана, варьируется объем используемого словаря и проч.). Кроме того, существует множество «смешанных» алгоритмов, где текст, сжатый, например, с помощью алгоритма PPM, дополнительно кодируется с помощью кода Хаффмана.

Все эти алгоритмы реализованы в различных программах, которых в настоящий момент существует довольно много. Каждая из них реализует разные варианты алгоритмов сжатия данных. Дополнительное разнообразие возникает из-за того, что у многих программ имеется несколько версий, которые также имеют разные алгоритмы сжатия. В работе [6] приведены некоторые результаты эксперимента по сравнению точности определения авторства текста с использованием алгоритмов сжатия данных.

Ряд экспериментов проводился на массиве новостей агентства «Рейтерс» (Reuters Corpus Volume 1). Было отобрано 50 авторов с наибольшим объемом статей, всего 1 813 статей. Выборка случайно была разбита на 10 равных частей, одна из которых использовалась для тестирования. Лучший результат был получен для метода с применением программы **rar** (точность 0,89).

Другой ряд экспериментов был проведен на корпусе текстов, состоящем из 385 текстов 82 писателей. Тексты подверглись предварительной обработке. Во-первых, были склеены все слова, разделенные переносом. Далее были выкинуты все слова, начинавшиеся с прописной буквы. Оставшиеся слова помещены в том порядке, в каком они находились в исходном тексте с разделителем из символа перевода строки. У каждого из писателей было отобрано по контрольному произведению. Остальные тексты были объединены в обучающие тексты. Объем каждого контрольного произведения составлял не менее 50 000–100 000 символов. Проведенные исследования показали, что программы сжатия угадывают истинных писателей весьма часто на текстах большого объема. Особенно хорошо проявляет себя программа **rarw** (точность 0,71), результаты применения которой превосходят реализацию других подходов в

этой области. Тем не менее остаются и открытые вопросы. Например, почему использование программы **rarw**, применяющей модификацию алгоритма LZ, на файлах большого объема опережает многие другие методы, также применяющие модификацию LZ.

Применение методов из теории вероятностей и математической статистики для атрибуции текстов

В некотором роде продолжение данного исследования нашло себя в работе [7]. Предлагаемый метод основан на учете статистики употребления пар элементов любой природы, идущих друг за другом в тексте (букв, морфем, словоформ и т. п.), т. е. на формальной математической модели последовательности букв (и любых других элементов) текста как реализации цепи Маркова. По тем произведениям автора, которые достоверно им созданы, вычислялась матрица переходных частот употребления пар элементов (букв, грамматических классов слов и т. п.). Она служила оценкой матрицы вероятности перехода из элемента в элемент. Для каждого автора строилась матрица переходных частот и оценивалась вероятность того, что именно он написал анонимный текст (или фрагмент текста). Автором анонимного текста считался тот, для кого вычисленная оценка вероятности больше.

Исходный корпус текстов в результате предварительной обработки был представлен в следующих вариантах:

а) пары букв в их естественных последовательностях в тексте – в словах (в той форме, в которой они употреблены в тексте) и пробелах между ними;

б) пары букв в последовательностях букв в приведенных (словарных, лемматизованных или исходных) формах слов; например, предыдущее предложение в таком случае предстает в виде «пара буква в последовательность буква приведенный словарный лемматизованный или исходный форма слово»;

в) пары наиболее обобщенных грамматических классов слов в их последовательностях в предложениях текста. К таким классам слов относят части речи (существительные, глаголы, прилагательные и т. п.) и некоторые условные категории вроде «конец предложения», «сокращение» и др.;

г) пары менее обобщенных грамматических классов слов. К ним относятся такие семантико-грамматические разряды, как одушевленные существительные, неодушевленные существительные, прилагательные качественные, относительные, притяжательные и т. п.

В процессе предварительной обработки отбрасывались все слова, для которых не удалось автоматически определить грамматический класс, все знаки препинания, *все слова с заглавной буквы*, склеивались все слова, разделенные переносом. Каждый символ кодировался числом.

Была произведена перекрестная проверка метода на материале 385 текстов 82 авторов. Показателем точности метода являлся процент правильно определенных произведений. Для варианта *а* получено 73 % точных определений, для *б* – 62 %, для *в* – 61 %. На материале варианта *г* получены существенно худшие результаты – 4 %.

В работе [8] показано, что последовательность символов текста не обладает свойствами простой цепи Маркова. Таким образом, гипотеза, выдвинутая в [5; 7], опровергнута. Тем не менее на основе проведенных в [7] экспериментов был сделан вывод, что использование пар подряд идущих в тексте букв дает более точные результаты, чем использование таких языковых категорий, как одиночные грамматические классы слов и их пары. Поэтому выдвинуто предположение, что в буквенных парных структурах частично отображаются полные структуры морфем словоформ текста – префиксальные, корневые, суффиксальные и флективные. Тем самым довольно большой объем словоизменительной и словообразовательной информации о структуре русских слов оказывается отображенным в статистике парной встречаемости букв, что и определяет высокий уровень эффективности использования этой статистики для определения авторства текста. Другими словами, подсчет частот употреблений пар букв по-

зволяет учесть информацию о словаре, который используется автором, а также, косвенно, информацию о предпочитаемых им грамматических конструкциях.

Система «Атрибутор»

Как продолжение развития подхода, использующего в качестве стилевых признаков бинарные буквосочетания, А. Н. Тимашев [2] предложил применять трехбуквенные сочетания – триады. При таком методе анализу поддаются однобуквенные и двухбуквенные служебные слова, а это значительная часть наиболее частотных предлогов, союзов, частиц и междометий, которые традиционно считаются значимыми стилеметрическими показателями. По этой причине двухбуквенные, четырех- и более буквенные цепочки менее показательны, что и было доказано в процессе исследования.

На основе данных рассуждений был создан программный продукт для автоматического сравнения и классификации текстов по параметрам индивидуального авторского стиля под названием «Атрибутор» (http://www.textology.ru/web.htm).

База этой программы содержит произведения 103 авторов и использует экспертную обработку текстов. В эталонную выборку, на которой происходило обучение «Атрибутора», попали в основном романы и повести отечественных писателей XIX–XX вв. Пополнение шло за счет ресурсов известных электронных библиотек, наибольшее количество текстов было получено в библиотеке М. Мошкова. Выборка подбиралась таким образом, чтобы тексты разных писателей в максимальной степени различались друг от друга, а тексты одного писателя были максимально близки. Те случаи, когда известный писатель в какой-то период своего творчества резко менял стиль изложения, отсеивались.

Для обработки текста «Атрибутором» необходимо, чтобы его длина была не меньше 6 страниц. Ограничение на длину текста накладывается для того, чтобы избежать ошибок, связанных со сравнением статистически несопоставимых объектов. В обработку попадают все слова текста, за исключением имен собственных.

Система «СМАЛТ»

Существуют подходы, основанные на изучении особенностей синтаксических структур текста: деревьев зависимостей и типов связей [9], деревьев зависимостей и мер сложности [10]. Кроме того, были проведены исследования, результатом которых явилась реализация системы атрибуции текстов «СМАЛТ» (Статистические методы анализа литературного текста), основанная на алгоритмах автоматизации морфологического и синтаксического анализа. Подробное описание системы можно найти в [2; 11].

Обработка текстов в этой системе производилась в несколько этапов. На первом шаге выполнялось автоматизированное разбиение исходного текста на лексические единицы, среди которых выделялась часть (или раздел), абзац, предложение, слово. На втором этапе осуществлялась автоматическая обработка текста и его морфологический разбор. На базе построенного морфологического разбора производилась третья стадия обработки текста – синтаксический анализ. Базой данных литературных произведений для проведения исследований явилась 81 публицистическая статья 60–70 гг. XIX в. журналов «Время» и «Эпоха», целью – установить, является ли действительным автором выбранных статей Ф. М. Достоевский.

В работе [11] была выдвинута гипотеза об эффективности выполнения некоторых методов для анализа текстов: метода проверки гипотез с помощью критерия Стьюдента, критерия Колмогорова – Смирнова на согласованность с заданным распределением, методов кластерного анализа, методики «сильный граф», в которой в качестве основной характеристики текстов рассматривалась матрица частот парной встречаемости грамматических классов слов.

Для проведения эксперимента с помощью критерия Стьюдента в качестве параметров были взяты следующие величины: средняя длина слова в буквах, средняя длина предложения в словах, индекс разнообразия лексики (отношения числа разных словоформ к числу словоупотреблений). Проводилось исследование с выборками разных объемов: в 200, 300, 400, 500

[2] *Тимашев А. Н.* Атрибутор // Текстология. ru. 2002. URL: http://www.textology.ru/atr_resum.html

и 600 слов. В итоге, были получены числовые значения критерия Стьюдента для всех статей. Среди группы статей Ф. М. Достоевского выявлялась статья с максимальным значением *t*-характеристики. Среди группы атрибутируемых статей и статей других авторов исключались статьи со значением *t*-характеристики, большим фиксированного.

При работе с такими параметрами, как общее распределение длины слова, общее распределение длины предложения, лексический спектр текста на уровне словаря и лексический спектр текста на уровне текста ставилась задача определения вероятности того, что распределения длин слов в буквах в двух статьях, одна из которых – объединение статей Ф. М. Достоевского, взяты из одной и той же «генеральной совокупности» и могут рассматриваться как управляемые одними и теми же закономерностями. Для этого использовался непараметрический критерий Колмогорова – Смирнова. Использовались частотные словари на каждые 500 слов текста. Все словоформы распределились в группы по 1, 2, …, 10 раз встречаемости в выборке. Далее определялось число словоформ в каждой группе, что означает распределение частот на уровне словаря, и покрываемость текста, что означает распределение частот на уровне текста.

В результате экспериментов не удалось установить, является ли автором рассматриваемых статей Ф. М. Достоевский, так как обе гипотезы (о том, что Ф. М. Достоевский – автор, и о том, что не автор) не верны. Причем была доказана независимость результатов исследования от видов текстов (авторская или современная орфография и пунктуация).

В методе иерархической кластеризации использовались 2 меры расстояния между объектами: Евклидова мера и мера Чебышева. Для определения расстояния между кластерами использовались методы ближнего и дальнего соседа. Исследование проводилось на основе двух наборов признаков: основного, состоящего их частей речи (16 признаков), и расширенного, с подключением дополнительных морфологических параметров, например падеж, род и т. п. (156 признаков). Применение методов корреляционных плеяд и иерархической кластеризации показало неэффективность использования формально-грамматических параметров для классификации исследуемых статей с целью решения задачи атрибуции, более того, было доказано, что увеличение числа этих параметров не улучшает результаты исследования.

Применение методики «сильный граф», основанной на изучении закономерностей расположения частей речи в рамках предложения по определенным параметрам, не позволило четко и однозначно ответить на вопрос о принадлежности ряда статей Ф. М. Достоевскому.

Еще один недостаток предложенных методов состоит в том, что задачу определения авторства приходится сводить к задаче построения качественного и быстрого синтаксического анализатора. Последняя из задач является не менее трудной и в настоящее время не решена на требуемом уровне.

Система «Антиплагиат»

Среди программных продуктов для определения авторства текстов можно выделить систему «Антиплагиат» (http://www.antiplagiat.ru). Этот интернет-сервис предлагает осуществить проверку текстовых документов на наличие заимствований из общедоступных сетевых источников. Система позволяет проводить атрибуцию текстов на различных языках.

На первом этапе система собирает информацию из различных источников: загружает из Интернета и обрабатывает сайты, находящиеся в открытом доступе, базы научных статей и рефератов. Загруженные документы проходят процедуру фильтрации, в результате которой отбрасывается бесполезная с точки зрения потенциального цитирования информация (например, HTML-страницы с большим количеством рекламы, новостные заголовки и т. д.).

На следующем этапе каждый из полученных таким образом текстов определенным образом форматируется и заносится в системную базу данных. Кроме того, в общую базу текстов поступают документы, загруженные на проверку пользователем, если такая возможность была разрешена им во время процедуры загрузки. Все пользовательские документы, загружаемые для проверки, ставятся в очередь на обработку.

Поиск совпадений осуществляется методом сравнения последовательностей символов без учета языковых особенностей и речевых взаимосвязей. За счет этого достигается высокая, в несколько секунд, скорость поиска совпадений. Проверка документа, например, реферата

среднего размера, занимает несколько секунд. После проверки документа пользователь получает отчет, в котором представляются результаты. Структура отчета позволяет выделять в проверяемом тексте заимствованные части как по всем источникам, так и по их любому подмножеству.

Все программные алгоритмы, используемые в «Антиплагиате», являются коммерческой тайной компании «Форексис», и открытого доступа к ним нет.

К недостаткам системы можно отнести невозможность «отлавливать» заимствованный текст при условии, что в каждом из предложений текста добавлено или убрано всего лишь одно слово. На данный момент существуют программы, например «Антиплагиат киллер» (http://otlichnik.biz/publ/antiplagiat_killer_2_0/1-1-0-4), позволяющие «обходить» систему «Антиплагиат».

Авторский инвариант и лингвистические спектры

В рамках относительно небольшого текста значения большинства формальных характеристик не позволяют установить авторский стиль. Кроме того, на коротких текстах часто не проявляются и другие характеристики, например, особенности использования авторской фразеологии и идиоматики, а также метафорической системы, системы эпитетов и т. д. С другой стороны, грамматические особенности авторского стиля – частота употребления неполнозначных, служебных слов (частиц, союзов, предлогов, некоторых модальных слов, вводных выражений) – для текстов порядка 1 000–2 000 слов сохраняются. Такой метод определения авторства текста иногда называют лингвостатистическим анализом неполнозначной лексики.

В своей работе В. П. Фоменко и Т. Г. Фоменко [4] вводят понятие *авторского инварианта* – формальной характеристики текста, удовлетворяющей условиям массовости, устойчивости и различающей способности. Авторский инвариант – характеристика текста, вычисленная как процент содержания служебных слов (союзов, предлогов, частиц – всего 55 слов) в тексте. Начальная выборка состояла из 2 000 слов, затем объем выборок последовательно увеличивался следующим образом: 4 000, 8 000, 16 000 слов. Проведенный эксперимент показал, что дальнейшее увеличение объема выборок необязательно, так как искомый авторский инвариант был обнаружен уже при величине выборки в 16 000 слов. В качестве критерия стабилизации был взят следующий принцип. Объем выборки увеличивался до тех пор, пока не обнаруживался параметр, для которого средняя величина его отклонений от средних значений вдоль произведений всех исследуемых писателей оказывалась существенно меньше амплитуды колебаний параметра между текстами разных авторов.

Эксперименты проводились на выборке из основных произведений 23 авторов XVIII–XX вв. В результате, например, был сделан вывод о том, что автором романа «Тихий Дон» не является М. А. Шолохов.

Серьезным ограничением этого метода является очень низкая разделительная способность оценки в случае большого числа авторов (потенциально метод может разделять лишь 10 авторских стилей).

Истоки этого метода восходят к работе [12]. Н. А. Морозов первым заметил, что именно особенности употребления служебных слов, лексем с общей семантикой, не привязанной к тематике художественного произведения, формируют авторский стиль и практически не поддаются имитации. В качестве характеристик авторского стиля он предложил брать часто используемые слова – предлоги, союзы, частицы, подсчитывая число употреблений каждой в отдельности. А графическое изображение их частот назвал *лингвистическими спектрами*. В конечном счете, выяснилось, что лингвистические спектры слишком неустойчивы, чтобы служить серьезным основанием для разграничения авторского стиля.

Система «Стилеанализатор»

Проблему атрибуции текстов в работах [8; 13] предлагается решать при помощи нейронных сетей и методов иерархической кластеризации. В качестве меры сравнения матриц частот появления признаков в исследовании использовалась мера Кульбака и мера хи-квадрат.

В работе также показано, что мера Хмелева из [5] является частным случаем меры Кульбака. В [8] предложены подходы для сравнения стилей текстов по частотным признакам с использованием гипергеометрического критерия (двустороннего точного критерия Фишера) и критерия хи-квадрат. Под частотным признаком понимается любой признак стиля текста, допускающий возможность нахождения частоты его появления в тексте (например, число появления абзацев в тексте). На основе проведенных исследований разработан программный комплекс «Стилеанализатор».

Проведены исследования зависимости от объемов текстовых фрагментов качества классификации текстов по авторству, по жанровым типам и источникам с помощью деревьев решений, метода Хмелева и метода с использованием нейронных сетей. В экспериментах было взято два набора текстов: художественных произведений (156 текстов, три подмножества: 30, 20 и 10 авторов) и газетных статьей (5 697 текстов, 57 журналистов за 2003–2004 гг.). Рассмотрены количественные признаки трех уровней: уровня букв, слов и предложений. Всего 14 различных наборов признаков.

Было обнаружено, что для разных текстов, с разным числом классов, для разных наборов признаков существует примерно постоянное минимальное значение объема фрагментов для приемлемой классификации. Оно составляет 30 000–40 000 символов, или 5 000–6 000 слов, или 400–600 предложений.

Использовались нейронные сети, обучающиеся без учителя и предназначенные для обработки больших массивов многомерной информации, – самоорганизующиеся карты Кохонена (Self-organizing map – SOM). За последние годы это направление является одним из наиболее развивающихся. С помощью SOM-сетей решаются многие проблемы классификации, обработки естественного языка, изображений, тестирования и обучения. Несмотря на широкое использование, SOM-сетям не хватает теоретической обоснованности – они опираются в основном на эмпирические результаты.

В итоге был получен вывод о том, что в случае удачного нахождения универсального набора характеристик можно обрабатывать любое число авторов и текстов (большие массивы информации). Достаточно постоянно модифицировать карту, добавляя новые произведения, и оценивать, как они взаимодействуют с ранее присутствующими.

Одним из серьезных недостатков метода является невозможность прогнозирования успешного результата. Генетический поиск на заданном наборе текстов может никогда не найти хороший вариант для разделения характеристик. Нет никакого критерия того, в правильном ли направлении движется поиск, верно ли он делает скачки, нужную ли скапливает информацию об исследуемом пространстве. Исследователь сам должен производить мониторинг поиска и следить за всеми «поворотами событий». Кроме того, нет механизмов, определяющих, сколько времени осталось до конца работы алгоритма, до того момента, когда дальнейший поиск не принесет своих результатов.

Другой проблемой метода является его трудоемкость. Число загруженных текстов, которое напрямую влияет на качество поиска, требует существенных ресурсов от вычислительной системы (большой объем памяти и мощный процессор). Для нахождения по-настоящему универсальных характеристик необходимо обработать массивные корпуса текстов, чтобы можно было с уверенностью заявить об их универсальности.

Проведенные эксперименты показали, что метод Хмелева и его модификации выигрывают как в скорости обучения, так и в качестве классификации. Нейронные сети дают сопоставимое качество, но сильно проигрывают в скорости. Деревья решений обеспечивают наихудшее качество классификации, но при этом дают наглядный вид решения и по ходу производят отбор самых информативных признаков.

Система «Авторовед»

Продолжение исследований по применению нейронных сетей в сочетании с методом опорных векторов при установлении авторства текстов нашло отражение в работе [1]. Если задачу определения авторства сформулировать как задачу классификации, то одним из широко применяемых выходов является построение бинарного классификатора. Все тексты, включая обучающую часть выборки, разворачиваются в очень большой вектор, индексируе-

мый словами. После этого имеется два множества точек из обучающей выборки в многомерном пространстве: принадлежащие данному автору и не принадлежащие автору. Для того чтобы разделить эти множества, нужно поделить пространство на две части. Самый простой способ сделать это – построить гиперплоскость. Такую гиперплоскость можно построить с помощью метода опорных векторов (SVM – Support Vector Machines). После этого для классификации текста с неизвестным автором достаточно проверить, в какую часть пространства он попал. Примерами применения метода опорных векторов при установлении авторства являются работы [1; 14].

Методы классификации с помощью SVM значительно превосходят многочисленных конкурентов: кластерную классификацию (когда документ соотносится к ближайшему множеству; в зависимости от определения расстояния до множества – имеется большое количество вариантов этого метода – средневзвешенный, *k* ближайших соседей и др.), а также наивную Байесовскую классификацию (которая предполагает, что частоты слов в тексте являются независимыми случайными величинами).

В качестве характерных признаков текста для описания авторского стиля предлагается брать наиболее частые триграммы символов и наиболее частые слова русского языка.

В качестве инструментов для атрибуции текстов в данной работе были выбраны искусственные нейронные сети архитектуры многослойный перцептрон (MLP), сети каскадной корреляции (CCN) и аппарат машины опорных векторов (SVM). CCN позволяют снизить временные затраты на обучение по сравнению с перцептроном за счет алгоритма автоматического построения топологии сети. SVM является наиболее точным из существующих в настоящее время методов классификации и в то же время наименее затратным по времени. Итоговое решение об авторе текста принимается ансамблем классификаторов по принципу мажоритарного голосования.

Основные результаты проведенных исследований были получены на корпусе, состоящем из 215 прозаических текстов 50 русских писателей. Тексты взяты из электронной библиотеки М. Мошкова. Размер каждого текста составлял более 100 000 символов. Использовались выборки объемом 10 00–100 000 символов (200–20 000 слов). Количество обучающих примеров каждого автора бралось равным 3, для тестирования использовалось по 1 выборке автора.

Эксперименты для случая 2, 5 и 10 авторов показали, что наиболее информативными авторским признаками являются ограничения в 300–700 наиболее частотных триграмм и 500 наиболее частых слов. Автора можно определить с точностью в среднем 0,95–0,98 при объеме текстовой выборки 20 000–25 000 символов. При этом начиная с 10 000 символов машина опорных векторов показывает лучшие из трех исследуемых классификаторов результаты. Установлено, что использование при идентификации автора комбинации частот букв русского языка, знаков пунктуации, наиболее частых триграмм символов и наиболее частых слов увеличивает точность идентификации в среднем на 0,06–0,12 на объемах текста до 10 000 символов.

Полученные методики были применены на практике для идентификации авторов коротких электронных сообщений во время внедрения разработанного метода и программного комплекса, названного «Авторовед», в деятельность воинской части 51952. Результаты показали, что авторство коротких текстов длиной 100 символов можно определить с точностью до 0,76 ± 0,11 в случае двух потенциальных авторов. При решении частной задачи по определению автора сообщения интернет-форума была достигнута точность 0,89 ± 0,08. Таким образом, предложенный метод дает довольно хорошие результаты на коротких электронных сообщениях, что выгодно отличает его от других ранее предложенных методов.

Заключение

В основе формальных методов атрибуции текстов лежит представление о том, что с возрастанием объема текста параметры, характеризующие авторский стиль, становятся устойчивыми с вероятностной точки зрения, что позволяет устанавливать авторство по стабильно повторяющимся формальным характеристикам текста. Поэтому более высокое качество атрибуции достигается для текстов большого объема, и менее точный результат получается для текстов маленького объема.

Сравнение программных средств атрибуции текстов

Название	Методы	Изменение параметров метода	Средства анализа текстов	Расширение перечня характеристик	Необходимый объем текста	Точность, %	Применение к решению реальных задач
Лингвоанализатор	Энтропийный подход, марковские цепи	Нет	Графем., стат. анализ	Нет	40 000-100 000 символов	84–89	Нет
Атрибутор	Марковские цепи	Нет	Стат. анализ	Нет	> 20 000 символов	Не изв.	Нет
СМАЛТ	Критерии Стьюдента, Колмогорова – Смирнова, кластерный анализ	Нет	Графем., морф., синт., стат. анализ, поддержка дореволюционной орфографии	Нет	500 слов для определения однородности	Не изв.	Да
Стилеанализатор	Марковские цепи, нейронные сети, деревья решений, меры расстояния	Да	Графем., стат. анализ, работа с размеченными текстами	Да	30 000-40 000 символов	90–98	Да
Авторовед	Нейронные сети, метод опорных векторов, QSUM	Да	Графем., морф., стат. анализ	Да	20 000-25 000 символов	95–98	Да
					100 символов	76	

Открытым остается вопрос о выборе авторского инварианта (набора формальных параметров текста). Часто на практике решается ограниченный круг задач для предварительно заданного набора текстов. Настройка, тестирование и демонстрация инструментов анализа ориентирована только на эти тексты, и нет никакой гарантии, что методы будут эффективно справляться с задачей на других данных. Иными словами, для построения универсального и независящего от текстов авторского инварианта необходимо искать новые пути формирования характеристик.

Установив набор характеристик, исследователь сталкивается с проблемой их структуризации, в чем существенную помощь могут оказать классические статистические методы. С помощью факторного анализа и анализа главных компонент можно установить вклад той или иной характеристики в процесс распознавания автора, иерархический кластерный анализ позволит сделать объединение отдельных характеристик в подгруппы, подгрупп в группы и т. д. Немалую помощь можно получить от нейронных сетей прямого распространения, если попытаться обучить сеть на наборе примеров, взяв в качестве входов отдельные характеристики, а затем оценивать, какое влияние оказывает тот или иной вход на систему выходов.

Недостаточно исследованы зависимости качества классификации различными методами от объемов фрагментов и от числа классов. Наконец, имеющиеся программы анализа текстов не ориентированы на комплексное исследование и сравнение стилей текстов (для разных задач анализа стилей текстов с использованием различных методов их решения, различных частотных признаков, различного текстового материала и т. д.). Наиболее удачное сравнение доступных программных средств для идентификации авторства текстов есть в [1] (см. таблицу).

К проблемам, затрудняющим исследования в области атрибуции текстов, относится также проблема составления выборки эталонных текстов. Желательно, чтобы произведения были подобраны следующим образом: тексты разных писателей в максимальной степени различались друг от друга, а тексты одного писателя были максимально близки. Но существует немало случаев, когда известный писатель в какой-то период своего творчества менял стиль изложения, или произведения были написаны в соавторстве. Эти факты создают дополнительные сложности при решении задачи установления авторства.

Необходимо проводить дальнейшие исследования, направленные на поиск новых или совершенствование уже имеющихся методов атрибуции текстов, а также на проведение экспериментов, целью которых является поиск характеристик, позволяющих четко разделять стили авторов, в том числе и на малых объемах выборки.

Список литературы

1. *Романов А. С.* Методика и программный комплекс для идентификации автора неизвестного текста: Автореф. дис. … канд. техн. наук. Томск, 2010. 26 с.

2. *Рогов А. А., Гурин Г. Б., Котов А. А., Сидоров Ю. В., Суровцова Т. Г.* Программный комплекс СМАЛТ // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды X Всерос. науч. конф. «RCDL'2008». Дубна, 2008. С. 155–160.

3. *Марков А. А.* Об одном применении статистического метода // Известия Императорской Академии наук. Сер. 6. 1916. Т. 10, № 4. С. 239–242.

4. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов // Новая хронология Греции: Античность в Средневековье. М.: МГУ, 1995. 422 с.

5. *Хмелёв Д. В.* Распознавание автора текста с использованием цепей А. А. Маркова // Вестн. МГУ. Сер. 9: Филология. 2000. № 2. С. 115–126.

6. *Хмелёв Д. В.* Классификация и разметка текстов с использованием методов сжатия данных // Все о сжатии данных, изображений и видео. 2003. URL: http://compression.ru/download/articles/classif/intro.html

7. *Кукушкина О. В., Поликарпов А. А., Хмелев Д. В.* Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. М.: Наука, 2001. Т. 37, № 2. С. 96–108.

8. *Шевелёв О. Г.* Разработка и исследование алгоритмов сравнения стилей текстовых произведений: Автореф. дис. … канд. техн. наук. Томск, 2006. 18 с.

9. *Севбо И. П.* Графическое представление синтаксических структур и стилистическая диагностика. Киев: Наук. дум., 1981. 192 с.

10. *Мартыненко Г. Я.* Основы стилеметрии. Л.: ЛГУ, 1988. 170 с.

11. *Рогов А. А., Сидоров Ю. В., Король А. В.* Автоматизированная система обработки и анализа литературных текстов СМАЛТ // Труды и материалы II Междунар. конгресса исследователей русского языка «Русский язык: исторические судьбы и современность». М: МГУ, 2004. С. 485–486.

12. *Морозов Н. А.* Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или другого известного автора. Стилеметрический этюд // Известия Отдела русского языка и словесности Императорской Академии наук. 1915. Т. 20, кн. 4. С. 93–127.

13. *Шевелёв О. Г.* Методы автоматической классификации текстов на естественном языке: Учеб. пособие. Томск: ТМЛ-Пресс, 2007. 144 с.

14. *Романов А. С., Мещеряков Р. В.* Идентификация автора текста с помощью аппарата опорных векторов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог-2009». М.: РГГУ, 2009. Вып. 8, № 15. С. 432–437.

Материал поступил в редколлегию 02.10.2012

T. V. Batura

FORMAL METHODS OF AUTHORSHIP ATTRIBUTION

This paper reviews the methods used for attribution of texts. The paper also provides a description of the popular software systems to determine the author's style, focused on the Russian language. An attempt was made to produce their comparative analysis, to identify features and drawbacks of approaches. The analysis of syntactic, lexical-phraseological and stylistic levels of text is the most interesting and the most difficult. Expert analysis of the author's style is a time consuming process, so the attention is paid to the formal methods of attribution. Currently, for establishing the authorship of texts following methods are used: the approaches of pattern recognition theory, methods of mathematical statistics and probability theory, neural network algorithms, cluster analysis algorithms, etc. Among the problems hampering research on attribution, the problem of choice of text parameters and sampling problem of reference texts are important. Further research is needed to find a new or improving of existing methods of text attribution, to search for characteristics that clearly separate styles of the authors, including short texts and small sample size.

Keywords: text attribution, authorship attribution, formal parameters of the text, author's style, classification of texts.