Е.С. Родионова, С.А. Хозяинов, О.А. Митрофанова

КОРПУСЫ ТЕКСТОВ В ИССЛЕДОВАНИЯХ ПО АТРИБУЦИИ ЛИТЕРАТУРНЫХ ПРОИЗВЕДЕНИЙ

0. Введение

В прикладных разработках в области математической лингвистики, при проведении экспериментальных исследований качество результатов в высокой степени зависит от того, насколько методы (лингвистические и математические) соответствуют языковому материалу и исследовательским целям. При атрибуции анонимных и псевдонимных текстов мера этого соответствия играет решающую роль, а состав корпусов текстов, формируемых для проверки атрибуционных гипотез, в значительной мере определяет ее результаты.

1. Предмет и метод атрибуционных исследований

Предметом исследований по атрибуции литературных произведений являются индивидуально-стилевые особенности текстов, приписываемых различным писателям. Наблюдения над развитием научной мысли в области параметризации авторских стилей позволяют заметить, что эффективный метод стилистического анализа в целях атрибуции предполагает применение многомерных классификаций, направленность на определение характеристик текста, а не отдельного предложения, на многоуровневое описание языка текста.

Этим требованиям отвечает методология атрибуции, основанная на идеях теории распознавания образов, которая уже нашла свое применение при решении целого ряда сложных проблем

авторства[1]. В 2004–2008 гг. на кафедре математической лингвистики СПбГУ с опорой на применение этой методологии были проведены два исследования: одно посвящено атрибуции стихотворных пьес, приписываемых Мольеру, другое – атрибуции публицистики, приписываемой А.С. Пушкину[2]. Интересна специфика материала этих исследований, определившая принципы формирования корпусов текстов и выбора методов их анализа.

2. Материал исследований

В терминах теории распознавания образов множество объектов, объединенных общими свойствами, называется «образ», «класс». В задачах атрибуции классы формируются на основе текстов, авторство которых заранее известно. Как правило, каждому классу соответствует корпус текстов одного автора. Согласно требованиям, выработанным в рамках квантитативно-структурного изучения авторских стилей, все исследуемые тексты должны быть когерентны в хронологическом, жанровом и, если необходимо, тематическом отношениях.

2.1. Проблема «Корнель – Мольер»

Изучение документально-исторических фактов и данных филологического анализа по данной проблеме позволило сформировать сложную атрибуционную гипотезу, описывающую возможность написания спорных пьес, приписываемых Мольеру, П. Корнелем, Ф. Кино и неизвестными авторами. При отборе атрибути-

---

[1] *Марусенко М. А.* Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Изд–во ЛГУ, 1990.

[2] *Родионова Е. С.* Лингвистические методы атрибуции и датировки литературных произведений (к проблеме «Корнель-Мольер»): Автореф. дис. …канд. филол. наук. СПб., 2008; *Хозяинов С. А.* Атрибуция публицистических произведений, приписываемых А.С. Пушкину (тексты 1830–1836 гг.): Автореф. дис. …канд. филол. наук. СПб., 2008.

руемых объектов и формировании априорных классов учитывалось требование соблюдения жанрово-стилевой однородности текстов. Анализируемые корпуса текстов составили только комедии в стихах.

2.2. Публицистика, приписываемая А.С. Пушкину

На авторство этой публицистики, помимо Пушкина, претендуют в основном П.А. Вяземский, А.А. Дельвиг, Н.В. Гоголь и О.М. Сомов. Внутренняя однородность каждого класса достигается за счет того, что входящие в него тексты отвечают четырем критериям: 1) подлинности авторства; 2) жанрового соответствия (обзоры литературы и рецензии, полемические статьи и заметки разнородного содержания); 3) хронологического соответствия (тексты 1830–1836 гг.); 4) законченности произведения (текст должен быть закончен и издан в этот период).

3. Получение и обработка данных

В современной корпусной и компьютерной лингвистике существует много практических задач, решаемых с использованием различных алгоритмов кластеризации, с применением методов машинного обучения и распознавания образов: это, прежде всего, классификация лексики в корпусах текстов, тематическая рубрикация документов для нужд информационного поиска и пр.

Данные процедуры можно выполнить при помощи стандартных пакетов программ, однако с алгоритмической точки зрения эти компьютерные средства рассчитаны на быструю и качественную обработку больших массивов структурированных данных, поэтому они не всегда позволяют учесть важную лингвистическую информацию об исследуемых объектах. В этом смысле, при анализе малых текстовых выборок и недостаточно структурированного текстового материала традиционные методы и инструменты вряд ли применимы. Качество результатов здесь напрямую зависит от тщательности лингвистической предобработки текстов

и оценки информативности параметров, характеризующих анализируемые тексты.

Поэтому на этапе лингвистического анализа текстов предпочтительна ручная обработка данных. Реализация алгоритма атрибуции, напротив, должна быть автоматической.

3.1. Структуризация текстов

Любому стилистическому исследованию художественного текста должна предшествовать процедура структуризации текста: выделение и подсчет цельных предложений в тексте, разбиение речевого материала на авторскую и чужую речь.

Обычно именно авторская речь составляет предмет атрибуции. Но в исследовании стихотворных пьес, приписываемых Мольеру, именно авторская речь – перечисление персонажей пьесы, описание декораций, действий, которыми сопровождаются реплики актеров, – не включалась в анализируемый материал. Чужая речь в пьесах оформляется двумя способами: 1) в виде реплик непосредственных участников разговора: CELIE Ah ! n'espérez jamais que mon coeur y consente[1]; 2) она может входить в состав таких реплик, и тогда заключается в кавычки: Puisqu'elle a sur mon coeur un pouvoir absolu, Il lui suffit de dire: «Ainsi je l'ai voulu»[2]. В текст, подлежащий дальнейшему изучению, была включена только чужая речь первого типа.

Для структуризации текстов, характеризующихся нормативным употреблением знаков пунктуации, очень хорошо подходит «формально-пунктуационный метод структуризации текста». В рамках этого метода предложение определяется как «последовательность символьных цепочек и пунктуационных знаков между ними от одного конечного знака до другого, где «символьная цепочка» – это текстовое графическое изображение словоупотреб-

[1] *Molière*. Sganarelle. I, 1.
[2] *Corneille*. La Galerie du Palais. VIII, 2.

ления, а «конечный знак» представляет собой составной пунктуационно-пространственный отрезок текста, позволяющий формально распознавать в строке ситуацию “конец предложения”»[1].

Стихотворные пьесы XVII века, составляющие проблему «Корнель – Мольер», характеризуются специфической пунктуацией и синтаксисом, поэтому формально-пунктуационный метод структуризации текста в данном случае оказывается недостаточно надежен. В роли «конечных знаков» здесь могут выступать такие пунктуационные знаки, как «.», «;», «!» и «?». Знаки «.» и «;» служат символами конца предложения вне зависимости от их правого окружения: они позволяют формально разграничивать предложения вне зависимости от их месторасположения в строке (середина или конец) и от того, с какой буквы (прописной или строчной) начинается следующий фрагмент текста. Знаки «!» и «?» безусловно фиксируют конец предложения тогда, когда они оказываются в конце строки. Все те случаи, когда они расположены в середине строки, определяются как неоднозначные – в анализируемых текстах за знаками «!» и «?» в середине строки следуют слова, начинающиеся со строчных букв. Для установления границ предложения в этом случае требуется не формальный, а смысловой подход. Знаки «!» и «?» не обозначают окончание предложения в том случае, если они выделяют одно – два междометия или эмоциональный оборот речи. Таким образом, были сформированы два априорных класса: Ω_1 (Corneille), мощностью 11 текстов и объемом 11103 предложений, и Ω_2 (Quinault), мощностью 3 текста и объемом 3125 предложений.

Иная ситуация с употреблением знаков препинания характеризует тексты, составляющие проблему авторства публицистики, приписываемой А.С. Пушкину. Здесь знаки конца предложения

[1] *Гринбаум О.Н.* Компьютерные аспекты стилеметрии // Прикладное языкознание: учебник / Отв. ред. А.С. Герд. СПб.: Изд-во С.-Петерб. ун-та, 1996. С. 456.

«!» и «?» употребляются в середине предложения в двух основных функциях: 1) разделяют несколько синтагм: а) первая – вопросительное предложение, вторая – повествовательное, содержащее ответ на вопрос, поставленный в первом; б) все синтагмы (две или более) – однотипные (восклицательные или вопросительные) предложения; в) синтагмы – разнотипные предложения, соотнесенные друг с другом каким-либо видом сочинительной или подчинительной связи («Нет ни осмотрительности, ни оглядки! перелистываешь книгу и изумляешься»[1]); 2) эмоционально или интонационно выделяют в предложении одно слово («На воле совесть его не мучила, нет, – первое чувство, родившееся в нем, была *любовь!* и его тревожила только одна мечта, что он *сирота!*»[2]).

В то же время знаки «!» и «?» в этих текстах нередко нормально завершают предложение. Поэтому в нетипичном употреблении этих знаков вполне можно видеть проявление воли автора. Дробление же таких предложений на отдельные цельные предложения, по существу, искажает авторскую речь. Принимая это положение, процесс разбиения текста на предложения можно полностью автоматизировать при помощи «формально-пунктуационного метода структуризации текста». Так была определена структура априорных классов мощностью k текстов, объемом n предложений: Ω_1 (Вяземский П.А.), k=18, n=946; Ω_2 (Гоголь Н.В.), k=29, n=613; Ω_3 (Дельвиг А.А.), k=36, n=591; Ω_4 (Пушкин А.С.), k=43, n=1001; Ω_5 (Сомов О.М.), k=9, n=296.

3.2. Параметризация текстов

Важнейшая задача атрибуции – фиксация стиля атрибутируемого произведения и сопоставление его со стилем предполагаемых авторов, – невыполнима без выявления специфических язы-

---

[1] *Дельвиг А.А.* Указ. соч. С. 266.

[2] *Дельвиг А.А.* Сочинения: стихотворения, статьи, письма / Сост., вступ. ст., коммент. В.Э. Вацуро. Л.: Худож. лит., 1986. С. 229.

ковых признаков текста. Инструментом описания таких признаков является априорный словарь параметров. В его основе лежат предложенные различными авторами языковые признаки (синтаксические и морфологические), отражающие наиболее существенные структурные особенности организации предложений.

Инвентарь параметров, составляющих исходное описание в исследовании проблемы «Корнель – Мольер», был составлен из 51 параметра, релевантного для описания текстов XVII века на французском языке. В исследовании публицистики, приписываемой А.С. Пушкину, использовались 49 исходных параметров, позволяющих описать русскоязычные тексты первой трети XIX в.

Правила параметризации текстов были разработаны с учетом специфики исследуемого материала. Параметризация текстов осуществлялась следующим образом.

Сначала на языке параметров из априорного словаря параметров были составлены предварительные описания априорных классов. Затем на основе этих описаний была произведена оценка эффективности параметров, направленная на отбор малого числа информативных параметров, наилучшим образом различающих априорные классы. Для свертки параметрических пространств большой размерности сегодня успешно применяется алгоритм сингулярной декомпозиции матриц (SVD)[1]. Однако при ограниченном наборе априорных параметров наиболее релевантные из них лучше отбирать по схеме М.М. Бонгарда, осуществляемой в два этапа: 1) определение параметров, релевантных для различения априорных классов; 2) свертывание параметрического пространства на подмножестве этих параметров[2].

Обычно получаемые таким образом наборы информативных параметров невелики. В эксперименте по атрибуции пьес, припи-

[1] SVD – singular value decomposition (**http://en.wikipedia.org/wiki/Singular_value_decomposition**).

[2] *Бонгард М.М.* Проблемы узнавания. М.: Наука, 1967.

сываемых Мольеру, такой набор составили пять параметров (X02 – число элементарных предложений, X04 – число сочиненных предложений, X21 – число спрягаемых форм глагола, X31 – число подлежащих, X32 – число местоимений-подлежащих), а в эксперименте по атрибуции публицистики, приписываемой Пушкину – четыре (X09 – число подчиненных предложений первой степени, X18 – число служебных слов, X24 – число наречий, X32 – число подлежащих). Шесть из этих девяти параметров относятся к синтаксическому строю предложения. Это подтверждает принимаемое многими исследователями предположение о том, что синтаксические особенности языка текстов обладают высоким стилеразличающим потенциалом.

На языке информативных параметров были определены координаты атрибутируемых объектов и априорных классов. Объемы выборок для объектов и априорных классов определялись при заданной относительной точности (стандартная ошибка оценки составила не более 0,05 величины оцениваемого признака)[1].

3.3. Классификация текстов

Используемая методика атрибуции характеризуется высокими показателями точности и полноты. Точность – вероятность того, что объект будет верно отождествлен с соответствующим образом – в обсуждаемых экспериментах составляет минимум 95%, тогда как в случае применения стандартных средств классификации / кластеризации этот показатель чаще всего оказывается ниже (ср. результаты тестирования лингвистических ресурсов на семинаре РОМИП[2]). Полнота – вероятность того, что объект будет идентифицирован в отношении какого-либо из образов – зависит от выбора алгоритма обработки данных. Так, при использовании

[1] *В поисках* потерянного автора: этюды атрибуции / М.А. Марусенко, Б.Л. Бессонов, Л.М. Богданова и др. СПб.: Филол. фак. С.-Петерб. гос. ун-та, 2001. С. 14–15.

[2] РОМИП (**http://romip.ru**).

хорошо зарекомендовавшей себя машины опорных векторов (SVM)[1] решается задача бинарного выбора (относится объект к классу или нет). Однако если классов несколько, то процедура распознавания усложняется, и может оказаться, что интересующий исследователей объект не будет охвачен классификацией.

Такой исход маловероятен при использовании упомянутой методики: полнота применяемого алгоритма составляет 100%, при этом возможны нечеткие решения о принадлежности объекта к тому или иному классу (в случае, например, соавторства).

В результате работы распознающего автомата в исследовании по атрибуции стихотворных пьес, приписываемых Мольеру, с очень большой степенью вероятности (больше 0,95) такие шедевры Мольера, как «Le dépit amoureux», «L'École des maris», «Les Fâcheux», «L'École des femmes», «Tartuffe», «Les Femmes savantes» оказались атрибутированы П. Корнелю. Еще четыре пьесы были атрибутированы П. Корнелю с вероятностью от 0,63 до 0,73. Драматургу Ф. Кино с вероятностью 0,68 оказалась атрибутирована комедия «L'Étourdi», а две пьесы – «Dom Garcie de Navarre» и «La Princesse d'Élide» – составили апостериорный класс, что указывает на возможное существование неизвестного автора.

В исследовании публицистики, приписываемой А.С. Пушкину, с вероятностью более 0,95 были атрибутированы семь статей (четыре из них – Пушкину, в том числе «Письмо к издателю», опубликованное в третьем томе «Современника» за 1836, и рецензия на «Невский альманах на 1830 год» Е. Аладьина, считающиеся произведениями поэта). Остальные тексты были атрибутированы с очень невысокой долей вероятности (от 0,211 до 0,578). В силу того, что приписываемые Пушкину тексты в среднем малы (из 46 таких текстов самый большой состоит из 85 предло-

[1] SVM – support vector machine (**http://en.wikipedia.org/wiki/Support_vector_machine**).

жений, самый малый – из двух), а предполагаемых авторов много (пять), перед нами встала необходимость оценки значимости таких решений.

3.4. Малые выборки и оценка нечетких решений

В области обработки новостных потоков (мониторинг новостных сообщений в электронных СМИ[1]), информационного поиска (выявление новостных текстов близкой тематики, обработки корпусов кратких аннотаций[2]) и пр. проблема малых выборок сейчас стоит наиболее остро. Основное препятствие – ограничение на использование статистического аппарата при недостаточности выборочных данных – до сих пор не преодолено, хотя обрабатываемый материал хорошо структурирован, и существуют вполне надежные автоматические инструменты лингвистического анализа текстов такого типа на разных уровнях. Поэтому лингвисты вынуждены решать стоящие перед ними за-

---

[1] См.: *Баглей С.Г., Антонов А.В., Мешков В.С., Суханов А.В.* Кластеризация документов с использованием метаинформации // Компьютерная лингвистика и интеллектуальные технологии: Материалы международной конференции «Диалог 2006» // URL: **http://www.dialog-21.ru/dialog2006/materials/html/Bagley.htm**); *Ландэ Д.В., Брайчевский С.М., Дармохвал А.Т., Морозов А.Ю.* Веб-пространство и материалы информационных агентств // Компьютерная лингвистика и интеллектуальные технологии: Материалы международной конференции «Диалог 2008» // URL: **http://www.dialog-21.ru/dialog2008/materials/html/46.htm** и др.

[2] См.: *Белов А.А., Волович М.М.* Автоматическое распознавание тематики сверхкоротких текстов // Компьютерная лингвистика и интеллектуальные технологии: Материалы международной конференции «Диалог 2007» // URL: **http://www.dialog-21.ru/dialog2007/materials/html/05.htm**; *Ingaramo D., Pinto D., Rosso P., Errecalde M.* Evaluation of internal validity measures in short-text corpora // Proc. 9th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing–2008, Springer-Verlag, LNCS(4919), pp. 555–567 // URL: **http://www.dsic.upv.es/ ~prosso/** и др.

дачи, прибегая к более трудоемким процедурам, предполагающим построение и дальнейшее использование модулей словарной поддержки и формальных онтологий.

Аппарат принятия решений, используемый в теории распознавания образов, очень гибок и потому позволяет решать классификационные задачи разной сложности. Применительно к проблеме установления авторства публицистики, приписываемой А.С. Пушкину, для оценки значимости нечетких решений классификации полезным оказалось следующее.

Пусть мы имеем матрицу взвешенных евклидовых расстояний между распознаваемыми объектами (текстами со спорным авторством) и классами (эталонами классификации), а на ее основе определены $m = i \times j$ значений $P(X_i \in \Omega_j)$, т. е. m значений вероятности принадлежности i-го объекта j-му классу. Пусть для объекта X_i вероятность $P(X_i \in \Omega_z)$ – наибольшая. Определим для нее $r = j - 1$ значений «оценки значимости вероятностей»[1], которую обозначим через k:

$$k_{P(X_i \in \Omega_z)} = \frac{P(X_i \in \Omega_z)}{P(X_i \in \Omega_j)}, j = \overline{1,n}, j \neq z, \quad (1)$$

где n – число классов. Получив набор из r таких оценок, вычислим их среднее, которое будет показывать, насколько существенно смещение координат распознаваемого объекта X_i к той области взвешенного евклидова пространства, которая соответсвует наиболее вероятному решению – $P(X_i \in \Omega_z)$:

$$\bar{k}_{P(X_i \in \Omega_z)} = \frac{1}{n-1} \sum \frac{P(X_i \in \Omega_z)}{P(X_i \in \Omega_j)}, j = \overline{1,n}, j \neq z. \quad (2)$$

Значение $\bar{k}$ тем выше единицы, чем определеннее принятое решение. Эту процедура, как и процесс построения матрицы расстояний и вычисления вероятностей, легко алгоритмизируется.

[1] *Якубайтис Т.А., Скляревич А.Н.* Вероятностная атрибуция типа текста по нескольким морфологическим признакам. Рига: ИЭВТ, 1982. С. 15.

При надобности можно ввести порог, который позволит автоматически фильтровать решения, разделяя их на два класса – удовлетворяющих и не удовлетворяющих некоторому условию, например, $\bar{k} > 2$ (для 10 решений классификации значение $\bar{k}$ оказалось в интервале от 2,071 до 6,109).

4. Заключение

Важно, что обработанные тексты – это тексты особых жанров (стихотворные и публицистические), определенных исторических периодов. Поэтому полученные результаты могут быть весьма полезны для разработчиков специализированных поэтических, публицистических диахронических корпусов текстов[1] – при подготовке программных средств для извлечения и обработки информации из таких корпусов. В частности, ряд исследовательских операций по обработке статистических данных может быть введен в состав корпус-менеджеров для таких корпусов, чтобы предоставить исследователям новые возможности обработки текстовых материалов. С этой точки зрения, полученные результаты представляют интерес для создателей электронных коллекций и библиотек[2], в которых должны предусматриваться сервисы лингвистического анализа[3], а также для разработчиков систем диагностики плагиата[4].

---

[1] Например, подкорпусы НКРЯ (**http://www.ruscorpora.ru**), электронные антологии «Русский сонет», «Русский рассказ XIX–XX вв.» и др.

[2] Например, электронная библиотека М. Мошкова (**http://lib.ru**), русская виртуальная библиотека (**http://www.rvb.ru**) и др.

[3] Например, инструменты «Атрибутор» (**http://www.textology.ru/web.htm**) и «Лингвоанализатор», (**http://www.rusf.ru/books/analysis/**), а также сервис лингвистического анализа лаборатории фантастики; (**http://www.fantlab.ru**) и др.

[4] Например, инструмент «Плагиат–Информ» (**http://www.plagiatinform.ru/**) и др.