А.Ю. Хомепко, Т.В. Романова

АЛГОРИТМ АВТОМАТИЧЕСКОЙ АТРИБУЦИИ ПИСЬМЕННОГО ТЕКСТА В ЛИНГВОКРИМИНАЛИСТИКЕ

В рамках исследования была сделана попытка поиска алгоритма, с одно стороны, основанного на интеграции интерпретационного подхода к анали языкового знака и методов математической статистики и математического м делирования, а с другой - поддающегося автоматизации.

Целью работы явилось определение, могут ли выбранные методы мат матической статистики и стилеметрического анализа успешно применяться судебном автороведении (определить автора текста чаще всего требуется в сп рах, связанных с нарушением авторских и смежных прав (ст. 146 УК РФ) можно ли на их основе создать универсальную безошибочную методику атр буции текста *любого объёма* и можно ли автоматизировать процесс атрибуци текста на современном этапе развития российской судебно-экспертной практ ки.

Соответственно, ставилась **задача** – разработать алгоритм, включающи как методы интерпретационного анализа, так и методы математической стат стики и теории вероятности и позволяющий как можно более точно в автом тическом режиме определять автора письменного текста.

Материалы и методика исследования. Материалом для исследован послужили художественные тексты заведомо известных авторов, поскольку ц лью работы является определение того, *релевантна ли методика* стилеметрич ского анализа для текстов различных стилей и объёмов:

1) Тестовая выборка (ТВ) – выборка, на основе которой строилась и ходная модель. ТВ представляла собой тексты С.Д. Довлатова, размещённые Национальном корпусе русского языка (URL: http://www.ruscorpora.ru/), за и ключением текста «Наши» (1983 г.) (ЭТ). Этот текст рассматривался как эксп риментальный, то есть текст. у которого якобы не определён автор. Таким о разом, объём ТВ – 330709 слов; 2) ЭТ (экспериментальный текст) – текст С. Довлатова «Наши» (1983 г.), текст, автор которого якобы неизвестен. Объём 21230 слов.

В основу методики анализа было положено исследование Е.С. Родион вой[1], где описываются методы стилеметрического анализа, квантитативно лингвистики и теории распознавания образов. Методика Е.С. Родионовой бы совмещена с методикой анализа языковой личности по Ю.Н. Караулову[2], мет

[1] Родионова Е.С. Лингвистические методы атрибуции и датировки литературных произведение проблеме «Мольер - Корнель»). Автореферат диссертации на соискание степени кандидата филол тических наук. 2008. URL: http://epir.ru/pragmat/projects/corneille/files/authorefcrat.pdf

[2] Караулов Ю.Н. Русский язык и языковая личность. М.: Наука, 1987

дикой квантитативного анализа незнаменательных и стилистически немаркированных лексем А.Н. Баранова[1] и некоторыми постулатами теории вероятности.

В конечном итоге был выработан алгоритм со следующими этапами:

I. Построение атрибуционных гипотез об авторстве спорных текстов (тех текстов, автора которых якобы неизвестны);

II. Анализ языковой личности (ЯЛ). Отбор параметров для будущей математической модели. Результатом стали 35 выявленных характеристик ЯЛ С.Д. Довлатова на трёх уровнях языковой личности: 1) *вербально-семантический* уровень: местоимения *«я»*, *«мы»*, *«ты»*, *«они»*. Эти характеристики выделены, исходя из наличия соотношения в прозе Довлатова проблемы субъекта говорения и субъекта действия, то есть автора и героя. Использование сочинительных союзов *«а», «и», «но»* в начале предложения; 2) *лингвокогнитивный* уровень: *а) «плохо», «хорошо»* - лексемы, маркирующие отношение к действительности. Эти лексемы отражают в том числе аксиологические оценки; *б) «тёмный», «белый», «светлый», «грусть», «грустный», «грустно», «молчание», «молчаливый»* - лексемы, имплицитно и эксплицитно маркирующие отношение к действительности и создающие её образ; *в) «город», «чемодан», «детство», «родина»* - лексемы, вербализующие значимые для Довлатова концептуальные сущности. Образ города присутствует во многих произведениях Довлатова, являясь символом определённого типа сознания. Чемодан является у Довлатова символом дороги, перемещения; 3) *мотивационный* уровень: *а)* экспликаторы модальности допущения, неуверенности: *ну, пусть, ладно, пожалуй, так, бы, видно; б)* экспликаторы модальности удивления: *неужели, разве, ах; в)* экспликаторы модальности ограничения: *только, лишь, почти; г)* экспликатор модальности возражения: *всё-таки.*

III. Квантитативные и стилеметрические преобразования данных, полученных в результате анализа ЯЛ: 1) Определение выборочных частот. Механический подсчёт того, сколько раз параметр реализуется в каждой выборке; 2) Определение средневыборочной частоты по формуле (1):

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1+\ldots+x_n)$ формула 1, - где x_i - i-й элемент выборки, n – объём выборки; 3) Определение отклонения выборочных частоты от средневыборочной частоты (среднеквадратического отклонения рассчитывается по формуле (2)):

$s = \sqrt{\frac{n}{n-1}\sigma^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ формула 2, - где σ^2 — дисперсия; x_i — i-й элемент выборки; n— объём выборки; $\bar{x}$ — среднее арифметическое выборки (средневыборочная частота); 4) Поиск вероятной ошибки в определении средней частоты по формуле (3) (для α – 0,2 и вероятности 0,8 при (n – 1) степеней свободы (35-1=34): **t = 1,3070**):

$L = \frac{t\sigma}{\sqrt{k}}$ формула 3, - где t - табличный коэффициент (t-критерий Стьюдента); σ- среднеквадратичное отклонение; k - объём выборки. Для ТВ ошибка со-

[1] Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие, 2001. URL: http://rudocs.exdat.com/docs/index-418941.html

ставляет 0,002272751. Для ЭТ - 0,008969957. Ошибка для ТВ и ЭТ незначительна, но тем не менее, она учитывается во матрицах данных; 5) Определение релевантных параметров для конечных моделей. Определяются по t-критерию Стьюдента (4). Уровень значимости α – 0,2. Критическое значение – в таблице пересечение уровня степеней свободы (количества параметров - 1) и вероятности 0,8.

$t = \frac{|\overline{x_1} - \overline{x_2}|}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}}$ формула 4, - где $\overline{x_1}$, x_2 - средние арифметические; σ_1^2, σ_2^2 - стандартное отклонение; n_1, n_2 - объёмы выборок.

По результатам исследования выделены следующие релевантные для моделей параметры: модель ТВ и ЭТ1: *грусть, сочинительный союз "но" в начале предложения, пусть, грустный, разве, ах, белый, неужели.* Релевантными для построения моделей в настоящей работе считаются параметры, числовые показатели которых наиболее близки к табличному значению t-критерия (1, 3070). Важно, что ни одно значение параметра не превысило значение t-статистики. Для настоящего исследования это означает, что полученные результаты будет иметь точность менее изначально заявленной, то есть менее 80%;

IV. Переход от реальных объектов к их математическим моделям, то есть описание выделенных в ходе предшествующего анализа параметров с помощью условной сигнатуры. Формирование матриц данных; внение двух моделей: модели текстов-образов, описывающей некоторые закономерности языковой личности заведомо известного автора, и модели спорного текста, описывающей некоторые закономерности языковой личности якобы неизвестного автора. Для сравнения моделей используется коэффициент корреляции между однородными параметрами (5). Этот коэффициент показывает, насколько близки две модели. Чем ближе значение этого коэффициента к 1, тем более близки модели в качественном отношении.

$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$ формула 5, - где

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ - средние значения выборок. Коэффициент корреляции между числовыми значениями матриц ТВ и ЭТ равен 0,783448911306154.

Как видим, коэффициент корреляции действительно чуть ниже 80%, тем не менее, даже уровень приблизительно в 78%, на наш взгляд, можно считать значительным.

Таким образом, подтвердилась следующая атрибуционная гипотеза:

- H_0 – автор ТВ и ЭТ – одно лицо, то есть автор **ТВ и ЭТ - С.Д. Довлатов** (по закону транзитивности: если автор ТВ – С.Д. Довлатов, а автор ЭТ и ТВ – одно лицо, то автор ЭТ – тоже С.Д. Довлатов).

Дальнейшие разработки в выбранной сфере позволили выработать ряд рекомендаций для улучшения работы алгоритма.

Так, для использования методики в условиях реальной действительности для текстов большого объёма можно дать следующие рекомендации:

- число параметров для идентификации автора по письменному речевому произведению должно быть не менее 45 – 50 единиц;
- параметры должны представлять собой обширные синтаксические и морфологические классы, например, большинство материальных репрезентантов субъективной модальности (вводные слова, модальные частицы, междометия, конструкции с имени тельным представления).
- отбор параметров должен происходить на основе более глубокого анализа языковой личности автора текста-образца, причём в большем объёма именно на мотивационном уровне (возможно, также на вербально-семантическом). Для применения указанной методики без внесения в саму методику изменений можно дать следующие рекомендации:
- объёмы текста-образца и спорного текста должны быть близки и не должны превышать 10 страниц печатного текста каждый;
- тексты должны быть схожи с точки зрения функциональных стилей;
- методику можно дополнить вычленением из двух текстов (эталонного текста, то есть сравнительного образца, и спорного текста), так называемых, базисно-инвариантных лексем.

Е.Ю. Салдаева, В.М. Бухаров

ПРОСОДИКА СПОНТАННОЙ РЕЧИ: МЕТОДЫ ФОРМИРОВАНИЯ ЭКСПЕРИМЕНТАЛЬНОГО КОРПУСА И ЕГО АНАЛИЗА

Настоящий этап развития лингвистики характеризуется повышенным вниманием к звуковой стороне речи с функциональной точки зрения. Просодика, обеспечивая интонационную выразительность, играет важную роль для осуществления коммуникативной функции, так как именно с помощью просодических средств мы передаем то, что не можем передать на сегментном уровне. Благодаря исследованиям в данной области стало возможным создание систем психофизического распознавания голоса (по характеристикам темпа, ритма, мелодического оформления речи). Такие системы широко внедряются в нашу жизнь. Например, для оценки качества работы операторов контакт-центров.[1] Совершенствование подобных систем требует дальнейших исследований в области просодики. А благодаря развитию корпусной лингвистики, начиная с 60х годов прошлого века, в нашем распоряжении появляются речевые корпуса как база для лингвистических исследований. Речевой корпус – это «структурированное множество речевых фрагментов, которое обеспечено программными средствами доступа к отдельным элементам корпуса».[2] Для одних целей оказывается достаточным использование уже существующих корпусов, когда для других задач необходимо создание отдельного экспериментального корпуса.

---

[1] Эти и другие разработки в области речевых технологий и мультимодальной биометрии принадлежат российской компанией ООО «Центр Речевых Технологий». Постоянными клиентами ЦРТ являются такие крупные компании как Сбербанк, РЖД, АВТОВАЗ, TELECOMEXPRESS, Ростелеком, МТС и др.

[2] О.Ф.Кривнова Речевые корпуса на новом технологическом витке: //Речевые технологии: [Электронный ресурс] С.13. URL: http://speechtechnology.ru/files/2-2008.pdf. (Дата обращения: 04.02.2014).