УДК 808.2-1/-8:004

А. М. Амиева, А. А. Крамаренко, В. В. Филимонов, А. А. Живодёров

МАШИННАЯ АТРИБУЦИЯ РУССКОЯЗЫЧНЫХ ТЕКСТОВ: ОБЗОР МЕТОДОВ

Амиева Анастасия Михайловна
amieva_nastya@mail.ru
Крамаренко Анна Александровна
lonelywolf1333@gmail.com
Филимонов Виктор Валентинович
fvv1408@list.ru
Живодёров Андрей Алексеевич
csl@cbibl.uran.ru
ФГАОУ ВО Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Россия, Екатеринбург

MACHINE ATTRIBUTION OF RUSSIAN-LANGUAGE TEXTS: A REVIEW OF METHODS

Amieva Anastasia Mikhailovna
Kramarenko Anna Aleksandrovna
Filimonov Viktor Valentinovich
Zhivodyorov Andrey Alekseevich
The Ural Federal University named after the first President of Russia B.N. Yeltsin, Russia, Yekaterinburg

Аннотация. *В статье рассматривается проблема атрибуции и классификации русскоязычных текстов. Приводятся характеристики различных подходов и обосновывается выбор. В рамках каждого подхода предлагается параметр для классификации.*

Abstract. *The paper considers the problem of attribution and classification of Russian-language texts. There are characteristics of different approaches. Authors explain choice of approaches and propose parameters for classification.*

Ключевые слова: *атрибуция текстов, статистика χ^2, закон больших чисел, модель случайных блужданий, коэффициент диффузии.*

Keywords: *attribution of texts, χ^2 statistics, the law of large numbers, the random walk model, coefficient of diffusion.*

В последнее время быстро развиваются компьютерные методы исследования в гуманитарных областях, в том числе в лингвистике. Это связано с широким распространением и доступностью вычислительных ресурсов, развитием программ и алгоритмов, удобством обработки значительных объёмов данных, а также с тем, что результаты таких исследований стали востребованными в повседневной жизни и практике, например, в поиске различных интернет-ресурсов и в машинном переводе.

Задача классификации текстов по жанрам, простая на уровне обыденных действий, при попытке формализации оказывается весьма нетривиальной. «Так, если «объяснение» компьютеру понятий рифмы и ритмической размерности еще можно себе представить, то анализ «музыкальности», образности и эстетического воздействия кажется задачей, превосходящей по сложности проблему компьютерного анализа смысла текстов» [1].

Исследования текста проводятся в двух направлениях: пространственном и лингвистическом. Пространственное направление предполагает измерения таких характеристик, как длина строки, интерлиньяж, размер шрифта, его рисунок и др. К этим исследованиям можно отнести исследования удобочитаемости, понимаемой как совокупность пространственных характеристик текста и их влияния на чтение и понимание текстовой информации (Тарасов Д.А. [2, 3, 4], Тягунов А.Г. [2, 3], Сергеев А.П. [2, 3, 4], Филимонов В.В. [4] и др.). Лингвистическое направление включает в себя исследование смыслонесущих единиц, таких как предложения, фразы, синтагмы, а также структурных особенностей.

В нашем исследовании мы ставим конечной целью построение математической теории структуры текста, поэтому нас в первую очередь интересуют именно структурные исследования языка.

На данном этапе работы мы решаем задачу разработки методики машинной атрибуции текстов, которая может быть использована для установления авторства и оценки юзабилити, определённого в стандартах ISO, а также в российском ГОСТе [5, 6, 7] как эффективность, результативность и удовлетворённость пользователя.

Свои исследования [8, 9, 10, 11] мы строим на следующих предпосылках:

- в тексте существуют скрытые структурные элементы, обнаружение которых возможно специальными методами;
- смысл — это конвенциальный феномен, поэтому он исключается из рассмотрения из соображений требования объективности исследования. Конвенциальность смысла заключается в том, что автор и читатель по-разному интерпретируют текст в силу различия собственных установок и жизненного опыта.

Исходя из этих предпосылок, мы выбрали два подхода к решению указанных задач. Первый опирается на методики частотного анализа, второй — на математическую модель случайных блужданий. Оба подхода пригодны для решения задачи поиска скрытых структур в тексте, так как они не связаны со смыслом самого текста, то есть являются объективными и позволяют избавиться от конвенциальных эффектов. Также они оба связаны с цифровой обработкой данных.

Все исследования проводились на материале из специально созданного «Корпуса текстов русского языка» (далее Корпус), который включает в себя на сегодняшний день около 1 300 текстов художественного, научного, социально-политического, административного, религиозного направлений, а также из газетных и журнальных публикаций. Каждое направление представлено в виде соответствующего подкорпуса. Для переводных текстов указано двойное авторство: автор первоначального текста и автор перевода [9].

В основе первого подхода лежит частотный анализ, который базируется на том, что текст включает в себя слова, а слова — буквы. Значимыми показателями текста являются повторяемость букв, пар букв (биграмм) и вообще m-грамм, совместимость букв друг с другом, чередование гласных и согласных и некоторые др. В наших работах [8, 9, 10] исследовалась повторяемость отдельных букв и их троек (триграмм).

В исследованиях [8, 9] были рассмотрены статистические закономерности букворазмещений в текстах и отработана методика исследования текстов при помощи статистики χ^2. Результатом стало распределение текстов Корпуса по нескольким интервалам, связанным с величиной χ^2 и прагматикой самих текстов. Оказалось возможным достаточно отчётливо выделить несколько кластеров, которые условно могут быть названы: поэзия, художественная проза, научный, социально-политический и административный. То есть, несмотря на то, что машина не ориентируется ни на смысл текста, ни на его название и анализирует только последовательность знаков, кластеры, полученные с использованием статистики χ^2, в основном совпали с подкорпусами Корпуса, выделенными экспертно. Таким образом, экспериментально подтвердилась адекватность метода статистики χ^2 для анализа русскоязычных текстов.

Однако границы между кластерами оказались нечёткими, т.е. существуют области, в которых присутствуют тексты, принадлежащие различным подкорпусам. По-видимому, кластеризация текстов по одному параметру не во всех случаях позволяет однозначно отнести текст к определённому жанру, в некоторых случаях можно говорить лишь о вероятной атрибуции текста.

Следующим этапом исследования [10] стал поиск дополнительных параметров атрибуции текста. Было выдвинуто предположение, что различия между величинами статистики χ^2 носят случайный характер и могут быть связанны с конечностью длины текста. Чем больше длина текста, тем меньше должно быть стандартное отклонение (SD) значений χ^2, и в случае «достаточно больших» текстов оно асимптотически стремится к нулю согласно закону больших чисел:

$$SD = \frac{c}{\sqrt{N}}, \qquad (1)$$

где N — размер выборки (в нашем случае — количество гласных букв в тексте), c — коэффициент пропорциональности.

Таким образом, была поставлена задача определить зависимость стандартного отклонения значений χ^2 от длины текста, вычислив значение коэффициента c в формуле (1). Очевидно, что коэффициент c не зависит от длины текста N и может быть связан с особенностями самого текста, т.е. являться его атрибутом или быть характерным для языка в целом.

В результате исследования выяснилось, что для большей части рассмотренных текстов значение c лежит в диапазоне от 1,5 до 5, резко выделяются тексты религиозного подкорпуса (c от 0,3 до 3) и административного подкорпуса (c от 5 до 38). Также следует отметить, что тексты по истории и философии, отнесённые к научному подкорпусу, имеют близкие значения коэффициентов c (от 3,2 до 3,9), в то время как значения χ^2 для этих текстов сильно различаются (от 0,08 до 0,17).

Второй подход, реализованный в исследовании [11], основан на использовании математической модели случайных блужданий. В качестве элемента атрибуции предлагается коэффициент пропорциональности в законе Эйнштейна, условно названный нами коэффициентом диффузии текста.

В модели случайных блужданий текст рассматривается как цепочка случайных событий — появлений очередной гласной буквы. Основное допущение модели состоит в том, что процесс полагается полностью случайным, то есть появление новой гласной не зависит от предыдущей. Любое случайное блуждание может быть описано законом Эйнштейна, который для двумерного случая выглядит следующим образом:

$$\bar{R}^2 = 4Dt, \qquad (2),$$

где R — смещение, D — коэффициент диффузии текста (аналогичен коэффициенту диффузии для физической системы), t — время (соответствует порядковому номеру буквы от начала текста).

Нами были рассчитаны коэффициенты диффузии для текстов, взятых из пяти подкорпусов Корпуса: художественная проза, научный, административный, публицистический и религиозный. Выяснилось, что значения коэффициентов диффузии для одной группы текстов, куда вошли произведения художественной прозы, научные работы и публицистика, лежат в одном диапазоне (от 60 до 100), а значения коэффициентов диффузии административных и религиозных текстов лежат за пределами указанного диапазона.

Доверительные интервалы величины D для первой группы текстов перекрываются, то есть можно сказать, что эти тексты неотличимы друг от друга по коэффициенту диффузии, и его величина отражает некие общие свойства этих текстов. Таким общим свойством, на наш взгляд, может являться «направленность» текстов, под которой мы понимаем реализованную в тексте форму коммуникации между автором и читателем. Таких форм может быть две: субъект-субъектная и субъект-объектная. Первая предполагает партнёрские отношения, в некотором смысле соавторство читателя, совместный поиск смыслов. Вторая — регулирующее воздействие субъекта на объект. В художественных, научных и публицистических текстах реализуется субъект-субъектная форма коммуникации. В административных и религиозных — субъект-объектная.

По итогам проведённых исследований можно сказать, что рассмотренные подходы могут быть использованы для разработки алгоритмов машинной атрибуции текстов без предварительной экспертной оценки, они позволяют атрибутировать текст без учёта его смысла.

Список литературы

1. Горбич Л.Г. Опыт различения поэтических и прозаических текстов на основе сравнения распределений биграмм гласных букв / Л.Г. Горбич, В.В. Филимонов, А.А. Живодёров // Количественные методы в искусствознании: материалы Международной научно-практической конференции, 20-22 сентября 2012 г., г. Екатеринбург, 2013. — С. 163–166.

2. Тарасов Д.А. Обоснование и вывод формулы скорости чтения, основанной на учёте пространственных характеристик текстовой информации / Д.А. Тарасов, А.Е, Ахметова, А.П. Сергеев, А.Г. Тягунов // Информационные технологии, телекоммуникации и системы управления: материалы Международной научно-практической конференции, 15 декабря 2015 г., г. Екатеринбург. / УрФУ имени первого Президента России Б.Н. Ельцина, Екатеринбург, 2015. — С. 140–146.

3. Тарасов Д.А. Учёт пространственных характеристик полосы набора текста в формуле чтения / Д.А. Тарасов, А.П. Сергеев, А.Г. Тягунов // Известия высших учебных заведений. Проблемы полиграфии и издательского дела. — 2014. — № 6. — С. 3–10.

4. Tarasov, D.A. Legibility of textbooks: a literature review / D.A. Tarasov, A.P. Sergeev, V.V. Filimonov // Procedia — Social and Behavioral Sciences. — 2015. — С. 1300–1308.

5. ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals *(VDTs) — Part 11* [Электронный ресурс]. — Режим доступа: http://www.iso.org/iso/catalogue_detail?csnumber=16883 (дата обращения 24.12.2015).

6. ГОСТ по юзабилити [Электронный ресурс]. — Режим доступа: https://habrahabr.ru/post/203308/ (дата обращения 24. 12. 2015).

7. ГОСТ Р ИСО 9 241-210-2012 Эргономика взаимодействия человек–система. Часть 210. Человеко-ориентированное проектирование. — Введ. 29.11.2012. — М.: Стандартинформ, 2013. — 36 с.

8. Филимонов В.В. Экспрессия и упорядоченность в письменной речи / В.В. Филимонов, А.А. Живодеров, Л.Г. Горбич // Известия УрФУ. Серия 1 Проблемы образования, науки и культуры. — 2012. — №3 (104). — С. 313–319.

9. Филимонов В.В. Кластеризация русскоязычных текстов с применением статистики χ^2 / В.В Филимонов, А.М. Амиева, А.П. Сергеев // Информационные технологии, телекоммуникации и системы управления: материалы Международной научно-практической конференции, 12–13 января 2016 г., г. Екатеринбург. / УрФУ имени первого Президента России Б.Н. Ельцина, Екатеринбург, 2016. — С. 164–174.

10. Филимонов В.В. Атрибутирование русскоязычных текстов с использованием закона больших чисел / В.В. Филимонов, А.М. Амиева, А.А. Живодёров, А.А. Крамаренко // Информационные технологии, телекоммуникации и системы управления: материалы Международной научно-практической конференции, 12–13 января 2017 г., г. Екатеринбург. / УрФУ имени первого Президента России Б.Н. Ельцина, Екатеринбург, 2017 — [в печати].

11. Крамаренко А.А. Применение модели случайных блужданий для описания русскоязычных текстов / А.А. Крамаренко, В.В. Филимонов, А.А. Живодёров, А.М. Амиева // Информационные технологии, телекоммуникации и системы управления: материалы Международной научно-практической конференции, 12–13 января 2017 г., г. Екатеринбург. / УрФУ имени первого Президента России Б.Н. Ельцина, Екатеринбург, 2017 — [в печати].

УДК 316.77:004

Д.А. Богданова

О НЕКОТОРЫХ ПОСЛЕДСТВИЯХ КРАЖИ ИДЕНТИЧНОСТИ

Богданова Диана Александровна
d.a.bogdanova@mail.ru
Институт образовательной информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, Россия, г. Москва

ON SOME OF THE CONSEQUENCES OF IDENTITY THEFT

Bogdanova Diana Aleksandrovna
Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Russia, Moscow

Аннотация. *Рассматривается последствие кражи идентичности в социальной сети, приводящая к созданию фальшивых профилей. Впоследствии эти профили используются для мошеннических действий- любовных афер, катфишинга, когда мошенник, вводя жертву в заблуждение, выманивает у нее крупную сумму денег.*

Abstract. *The consequences of identity theft in the social network, leading to the creation of fake profile are considered. Subsequently, these profiles are used for fraudulent contravene the romance scams, catfishing, when the fraudster (catfisher), introducing the victim to misleading entices her a large sum of money.*