Ретивина Вероника Викторовна, Пакина Татьяна Александровна

ВОЗМОЖНОСТИ АТРИБУЦИИ ТЕКСТОВ НА ОСНОВЕ ТЕОРЕТИКО-ИНФОРМАЦИОННОГО ПОДХОДА

В статье рассматривается атрибуция текстов как задача распознавания образов. В качестве решающего правила классификации предлагается использовать критерий минимума информационного рассогласования. Приводятся экспериментальные данные статистического анализа художественных произведений, иллюстрирующие принцип его работы. Подробно излагается алгоритм распознавания для автоматизированной системы атрибуции текстов с применением теоретико-информационного подхода.

Адрес статьи: www.gramota.net/materials/2/2017/1-2/49.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2017. № 1(67): в 2-х ч. Ч. 2. С. 172-174. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2017/1-2/

УДК 81'33

В статье рассматривается атрибуция текстов как задача распознавания образов. В качестве решающего правила классификации предлагается использовать критерий минимума информационного рассогласования. Приводятся экспериментальные данные статистического анализа художественных произведений, иллюстрирующие принцип его работы. Подробно излагается алгоритм распознавания для автоматизированной системы атрибуции текстов с применением теоретико-информационного подхода.

Ключевые слова и фразы: атрибуция текстов; статистическая классификация; распознавание образов; информативные параметры; метрика; решающее правило; критерий минимума информационного рассогласования.

Ретивина Вероника Викторовна
Нижегородский государственный лингвистический университет имени Н. А. Добролюбова
retivina@mail.ru

Пакина Татьяна Александровна
Нижегородский государственный педагогический университет имени Козьмы Минина
ta_pakina@mail.ru

ВОЗМОЖНОСТИ АТРИБУЦИИ ТЕКСТОВ НА ОСНОВЕ ТЕОРЕТИКО-ИНФОРМАЦИОННОГО ПОДХОДА

Одной из актуальных задач литературоведения была и остается проблема атрибуции текстов. Основная ее цель – определение авторства литературного текста, а также установление жанра или времени его написания. В настоящее время накоплено большое количество методов и приемов атрибуции, в том числе основанных на формально-количественном подходе, с помощью которого процессу атрибуции придается более объективный характер.

Текст является продуктом языковой деятельности. Каждый текст отражает стилистические особенности своего источника. Существование точных количественных методов идентификации и проведение экспертизы на их основе могли бы разрешить большинство спорных вопросов в области атрибуции текстов. Поэтому одним из самых основных и самых сложных вопросов лингвистической статистики является выявление особенностей различных стилей (это могут быть стили авторские или стили функциональные) и их разграничение. В большинстве лингвостатистических исследований эти различия изучались с помощью количественного анализа функционирования некоторых языковых единиц в разных стилях. Однако результаты такого подхода не дают четкого ответа при решении задач атрибуции текстов, а, как правило, лишь определяют некоторую вероятность принадлежности исследуемого языкового материала к тому или иному стилю.

Целью данной работы является рассмотрение возможностей принципиально нового подхода к решению задач атрибуции. В предлагаемом методе задача атрибуции рассматривается как задача статистической классификации объектов на основе теоретико-информационного подхода. В отличие от традиционных методов сравнения количественных показателей различных стилей, новый способ позволяет получить четкий ответ на вопрос о принадлежности или непринадлежности изучаемого текста к одному из имеющихся классов.

В 1990 году была опубликована монография М. А. Марусенко [2], в которой проблема установления авторства текста впервые решалась методами распознавания образов на основе индивидуальных характеристик авторского стиля. В данной работе текст рассматривается как сложный лингвистический объект, для атрибуции которого используется многомерный статистический анализ, представленный в наиболее развитой форме – теории распознавания образов. Распознавание образов – это задача идентификации объекта по его характеристикам. Образ (класс) – классификационная группировка в системе классификации, объединяющая определенную группу объектов по некоторому признаку (применительно к текстам класс, например, образуют произведения одного автора или одного жанра). Пусть существует заранее заданное множество образов. Тогда задачей распознавания становится отнесение изучаемого объекта к одному из них на основе определенной методики, называемой решающим правилом.

В терминах распознавания образов стиль определяется как «набор свойств (параметров), характеризующих состав, способы объединения и статистико-вероятностные закономерности употребления речевых средств, образующих данную разновидность языка» [Там же, с. 17]. Этим набором свойств является совокупность информативных параметров, по которым проводится процедура распознавания. Помимо набора информативных параметров, необходимо выбрать метрику – способ определения расстояния между элементами. Чем меньше это расстояние, тем более похожими являются объекты. От выбора модели представления образов и реализации метрики зависит эффективность программы распознавания.

Для автоматической атрибуции текстов принцип решения задачи распознавания образов реализуется как алгоритм на основе метода множества эталонов. На входе его имеется набор обучающих выборок, по которым восстанавливаются статистические распределения, характеризующие поведение изучаемой совокупности параметров для каждого имеющегося образа A_i, метрика ρ и сам распознаваемый объект X. С помощью

метрики вычисляем расстояние $\rho(X, A_i)$ от X до каждого образа A_i. Объект X будет отнесен к образу, который окажется ближе всех, т.е. которому будет соответствовать наименьшее из вычисленных расстояний.

В Нижегородском государственном лингвистическом университете был разработан новый способ решения задач атрибуции, основанный на синтезе аппарата распознавания образов и теоретико-информационного подхода. В качестве решающего правила предлагаемый метод использует критерий минимума информационного рассогласования (МИР), который является интерпретацией классического критерия максимального правдоподобия в задачах распознавания дискретных объектов и имеет ряд преимуществ [4, с. 18]. В рамках нового критерия был разработан алгоритм статистической обработки текста, который позволяет осуществлять идентификацию текстов, причем этот алгоритм инвариантен по отношению к национальному происхождению языка.

В отличие от известных методов статистической обработки и атрибуции текстов, новый алгоритм выявляет различие стилевых особенностей, основываясь на величине информационного рассогласования, предложенного С. Кульбаком [1, с. 313]. Это рассогласование является мерой «расстояния» между двумя текстами и вычисляется по формуле:

$$\rho = \sum_{i=1}^{N} p_i \log(p_i / q_i) ,$$

где p_i и q_i – это вероятности появления i-го состояния изучаемого параметра (среди N возможных) в первом и втором текстах соответственно.

Критерий МИР справедлив для задач любой размерности, т.е. применительно к лингвистическим задачам это означает, что можно рассматривать поведение не какого-то одного признака в различных текстах, а поведение нескольких признаков в совокупности. Иными словами, данный метод позволяет изучать стиль как систему, учитывая все доступные лингвостатистические параметры одновременно.

В качестве иллюстрации работы алгоритма, основанного на применении двумерного критерия МИР, можно привести результаты следующего эксперимента по установлению авторства. Были найдены информационные расстояния между выборкой из романа Ф. М. Достоевского «Братья Карамазовы» и выборками из 10 других текстов (в том числе 5 других писателей). Все выборки были взяты в объеме 300000 символов. В качестве показателей их стилистических особенностей были выбраны длина предложения и количество запятых в каждом отдельно взятом предложении, т.к. «именно роль предложения как единицы языка в статистической картине организации стиля можно считать особой, точнее, особенно важной» [2, с. 53].

Для оценок вероятностей проводилась процедура вычисления относительной частоты появления в текстах соответствующих событий со стандартной регуляризацией.

В результате были получены следующие данные (см. Таблицу).

Таблица

Произведения	Рассогласование по числу запятых в предложении	Рассогласование по длине предложения	Рассогласование по двум признакам в совокупности
Достоевский – «Братья Карамазовы» Достоевский – «Неточка Незванова»	0,0085176550	0,0454579195	0,0894776095
Достоевский – «Братья Карамазовы» Достоевский – «Идиот»	0,0112268096	0,0575987123	0,1092604796
Достоевский – «Братья Карамазовы» Лесков – «Островитяне»	0,0088514073	0,0633261627	0,1035792699
Достоевский – «Братья Карамазовы» Лесков – «Соборяне»	0,0289878162	0,0775171403	0,1316942659
Достоевский – «Братья Карамазовы» Тургенев – «Отцы и дети»	0,0673546471	0,1252185633	0,1441897819
Достоевский – «Братья Карамазовы» Тургенев – «Новь»	0,0791181653	0,1263077986	0,1589098295
Достоевский – «Братья Карамазовы» Горький – «Дело Артамоновых»	0,0268260718	0,1006169866	0,2016652220
Достоевский – «Братья Карамазовы» Горький – «Фома Гордеев»	0,0391878138	0,1664174757	0,2441758933

Как видно из таблицы, для всех трех случаев наименьшим информационным рассогласованием с выборкой из романа Ф. М. Достоевского «Братья Карамазовы» обладают выборки из произведений того же автора. Поэтому естественно предположить, что для установления авторства анонимного произведения нужно сопоставить между собой информационные рассогласования между этим текстом и произведениями предположительных авторов. В соответствии с критерием МИР решение принимается в пользу автора, произведение которого составило минимальную величину информационного рассогласования среди всех альтернативных вариантов.

Следует также отметить, что величина информационного рассогласования для двумерного случая является более качественным показателем различий между текстами, чем рассогласование по любому из признаков в отдельности. На практике не удается выделить какой-либо один универсальный параметр для всех текстов, который бы четко иллюстрировал различия между ними. В частности, из таблицы видно, что рассогласование по количеству запятых в предложении и рассогласование по длине предложения менее четко разграничивают авторские стили, чем это делает величина информационного рассогласования, вычисляемая по этим двум признакам в совокупности.

Если в качестве идентификационных признаков текста рассматривать некоторую совокупность независимых стилевых параметров, принцип МИР сводится к проверке минимальности суммы информационных рассогласований по каждому из них [3, с. 26]. В этом случае решение задачи статистической классификации объектов с помощью критерия минимума информационного рассогласования значительно упрощается и сводится к вычислению суммы информационных рассогласований, полученных по каждому из параметров в отдельности. Таким образом, при реализации алгоритма распознавания решается проблема малости выборок и значительно снижаются вычислительные затраты.

В частности, при разработке автоматизированной системы атрибуции текстов на основе данного критерия необходимо:

1) На первом шаге применить алгоритм формирования набора информативных параметров, по которым будет производиться сравнение изучаемого текста с набором эталонных текстов, имеющихся в базе данных. При этом все измеряемые параметры разбиваются на m групп таким образом, что в каждую из них попадают параметры, сильно коррелированные между собой, а слабо связанные параметры попадают в разные группы. Затем в каждой группе находится наиболее информативный параметр. Этот параметр и будет включаться в конечный набор признаков (по одному из каждой группы), по которому будет производиться распознавание. Таким образом, происходит «свертывание» всего исследуемого набора параметров до m наиболее информативных параметров, слабо коррелированных между собой.

2) На втором шаге вычислить информационные рассогласования по каждому из m параметров между исследуемым текстом и каждым из эталонных текстов, а затем просуммировать полученные значения для каждого эталона.

3) Решение задачи установления авторства принять в пользу того эталона, которому будет соответствовать минимальная сумма информационных параметрических рассогласований.

Список литературы

1. **Кульбак С.** Теория информации и статистика. М.: Наука, 1967. 408 с.
2. **Марусенко М. А.** Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов. Л.: Изд-во Ленинградского университета, 1990. 168 с.
3. **Ретивина В. В.** Об одном подходе к разработке автоматизированной системы авторизации текстов // Ползуновский альманах. 2007. № 3. С. 25-27.
4. **Савченко В. В., Савченко А. В.** Принцип минимального информационного рассогласования в задаче распознавания дискретных объектов // Известия высших учебных заведений России. Радиоэлектроника. 2005. Вып. 3. С. 10-18.

THE POSSIBILITIES OF TEXTS ATTRIBUTION BASED ON INFORMATION-THEORETICAL APPROACH

Retivina Veronika Viktorovna
Nizhny Novgorod State Linguistic University named after N. A. Dobrolyubov
retivina@mail.ru

Pakina Tat'yana Aleksandrovna
Minin University
ta_pakina@mail.ru

The article examines the attribution of texts as the problem of image recognition. The authors propose to use the criterion of minimum of information discrepancy as the decisive rule of classification. Experimental data of statistical analysis of fiction illustrating the principle of its work are provided. The algorithm of the recognition for the automated system of attribution of texts with the use of information-theoretical approach is described in detail.

Key words and phrases: attribution of texts; statistic classification; image recognition; informative parameters; metrics; decisive rule; criterion of minimum of information discrepancy.