Технология атрибуции газетных статей: Анонимные журналисты «Пермских губернских ведомостей» начала XX века.

Сметанин Андрей Владимирович
студент
Пермский государственный университет, историко-политологический факультет Пермь, Россия
E-mail: andr_smetanin@mail.ru

Вопрос атрибуции текстов является постоянной проблематикой историков и филологов, поэтому в своей работе мы могли опираться как на опыт предшествующих исследователей, так и на собственные идеи.

Значительным достижением в данном вопросе стало привлечение математических средств к решению проблем авторства. В России это направление активно развивается с начала 1980-х гг., в связи с возможностью использования компьютерных технологий. Среди наиболее значительных вех можно назвать работы Милова Л.В. и Бородкина Л.И. с применением «графов сильных связей» (Милов, 1994). Работы их последователей – Злобина Е.В., Быстрова А.В., решавших проблемы авторства «Записок» И.И. Горбачевского (Злобин, 1990), предсмертного письма Б.В. Савинкова (Быстров, 1994). В последующие годы появилось множество других интересных методик атрибуции текста, которые нашли отражение в работах историков, филологов и математиков: «диаграмная энтропия» для анализа Синопсиса XVII в. (Тарнопольская, 1998), комплексное исследование текстов записей солдатских разговоров (Поршнева, 2002) и др. В настоящее время сильные школы функционируют в Москве, Санкт-Петербурге, Петрозаводске, Краснодаре.

В нашем случае, атрибутируемым источником является статейный материал газеты «Пермские губернские ведомости»[1], издававшейся с 1838 года и являвшейся единственной газетой дореволюционной Перми. Под этим общим названием выходила еженедельная официальная часть с постановлениями губернских властей, а также ежедневная неофициальная часть. Последняя имела подзаголовок «литературная, политическая и экономическая газета». Объектом нашего исследования является именно неофициальная часть.

В 2006 году Лабораторией исторической и политической информатики ПГУ была создана информационная поисковая система по номерам за 1909-1911 гг. (поддержана грантом РГНФ). Исследователю открылась возможность полноценного доступа к электронной версии газеты. Данный факт определяет выбор анализируемого периода.

Ключевой проблемой, обусловившей необходимость исследования, является слабая изученность дореволюционной журналистики в Перми, за исключением отдела литературной критики (Масальцева, 2006). Тому есть объективные причины, например, число анонимных статей в каждом номере составляет примерно 2/3 всех публикуемых материалов. Но даже среди статей и заметок, имеющих подпись, около половины отмечено псевдонимами (Чусовлянин, Синьор, Novu и т.п.). Псевдонимы повторяются весьма редко.

Перед исследованием стоят следующие задачи:

1) выработка или выбор технологии атрибуции текста

2) восстановление авторства статей, т.е. определение статей написанных одним человеком

3) воссоздание социо-культурного портрета пермских журналистов (проектная)

Анализируемый корпус имеет ряд особенностей, которые определяют выбор методики анализа. Ниже приведены основные из этих особенностей.

---

[1] Далее «ПГВ»

1) Небольшая величина статей. Самые большие из них достигают размера в 2 тысячи слов и не более трети газетной полосы, но такие статьи единичны. Средний размер колеблется в области 500 слов. В исследовательской практике, небольшой набор слов определяет возможность большей статистической погрешности. Важно отметить, что для анализа отобраны только авторские статьи, т.о. небольшие заметки формата «хроники» не учитывались.

2) Публицистический стиль. Главным минусом этого стиля, особенно в провинциальной дореволюционной прессе, является обилие штампов и цитат. Общеупотребительные шаблоны сужают пространство авторского стиля, что усложняет проблему определения авторства.

3) Отсутствие авторских эталонов. Нам неизвестно ни общее количество авторов, ни особенности стиля кого-либо из них, поэтому мы должны опираться на технологию, не требующую априорных знаний об авторе или корпуса его атрибутированных текстов.

4) Большое число прямых и скрытых перепечаток из центральных изданий. Часть материалов, выходивших в «ПГВ», имела явно неместное происхождение, например, рассказывалось в деталях о столичных событиях, но при этом указание на источник информации отсутствовало.

5) Проблемы с сохранностью источника. Из-за плохого качества бумаги, часть страниц была повреждена, и фрагменты текста утеряны.

Очевидно, что ключевой проблемой анализа является слабая выраженность авторского стиля (из-за специфического жанра и небольших размеров текста). Суммировав все требования к технологии атрибуции газетных статей, мы выработали следующие принципы:

1) Создание «условных авторских эталонов». Изначально выбирается корпус максимально больших статей, которые, однозначно, написаны местными журналистами (упоминаются местные реалии). После количественного анализа, наиболее близкие из них, которые можно отнести к перу одного автора, объединяются в один текст. Здесь работает логика: чем больше размер статьи, тем большее представление об авторском стиле мы получаем. Создание подобного «эталона» оберегает нас от перепечатанных статей непермских журналистов и оптимизирует дальнейшую работу. Уже на этом этапе можно получить определённое представление о количестве авторов.

2) В течение всего исследования сохраняется принцип постатейного анализа, так как мы исходим из нулевой гипотезы, что все статьи могут быть написаны разными авторами. Эта гипотеза помогает избежать преждевременных обобщений.

3) Каждая проанализированная статья сверяется на сходство с «эталонами».

Для установления авторства, математическому анализу подвергнуты следующие характеристики текстов: длина предложений, употребляемость частей речи, а также статистика парной встречаемости частей речи (разработка «графов сильных связей»). Эта методика, в сущности, близка к той, что использована Поршневой и Поршневым (Поршнева, 2002).

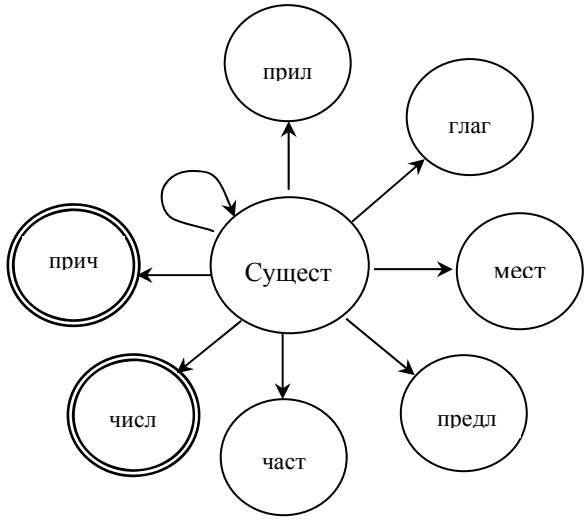
Методика, изложенная выше, базируется на синтаксическом анализе текста. В качестве дополнения мы вводим элемент лексического анализа. Из текста статей выделяются редко используемые или оригинальные слова, которые характерны для авторского лексикона. Составив своеобразное портфолио, в последствие мы можем с помощью запросов контент-анализа, проверять наличие таких лексических маркеров в ещё неатрибутированных текстах.

Для хранения кодировок текстов и упрощения расчётов, в работе мы использовали электронные таблицы программного пакета «Statistica V6.0».

Ниже приведён сокращённый пример использованного нами анализа. Имеется текст трёх статей (ПГВ, 1909): «В кулуарах» (за подписью N), «Отцы-юмористы» (подпись – С.И.), «Продажа людей с молотка» (без подписи). Специально подобраны статьи, выдержанные в ироническом тоне, похожие стилистически, по крайней мере,

внешне. К тому же, все материалы датированы мартом 1909 года, т.е. близки хронологически.

	№1.В кулуарах	№2.Отцы-юмористы	№3.Продажа людей с молотка
Кол-во слов/предложений	535/17	577/62	437/42
Наиболее весомые грамматические классы	существ – 31% глаголы – 14,6% местоим – 12% прилаг – 10,5% предлог – 9,3% союз – 8,6%	существ – 29,3% глаголы – 15,8% прилаг – 12,3% местоим – 10,2% союз – 8,7% предлог – 8,3%	существ – 38% глаголы – 14,6% предлог – 12,1% прилаг – 8% местоим – 8% наречие – 3,7%
Примечательные лексемы	Кулуар, Васька (*пройдоха*)	Сообразоваться, юмористика, поковыривать	Панель (*тротуар*), нумер



Тексты статей кодируются, для последующего оформления матрицы парной встречаемости грамматических классов, в соответствие в алгоритмом, предложенным Бородкиным Л.И. (Милов, 1994). После создания матрицы и исключения всех незначительных грамматических классов, имеющих относительную долю менее 1% от общего объёма текста результаты визуализируются в виде «графов сильных связей». Например, у «существительного» графы имеют схожий вид для всех трёх текстов, за исключением двух связей (с причастием и числительным), характерных только для статьи №3.

Небезынтересно сравнение относительных долей связей к общему числу всех последовательных связей в тексте.

	В кулуарах	Отцы-юмористы	Продажа людей с молотка
Наиболее весомые последовательности	Прил-глаг (8,7%) Сущ-сущ (8,5%) Сущ-глаг (5,4%) Мест-сущ (4,2%) Предл-сущ (3,5%) Глаг-сущ (3,5%)	Прил-глаг (9,1%) Сущ-сущ (8,3%) Сущ-глаг (4,5%) Предл-сущ (4,3%) Сущ-предл (3,7%) Прил-прил (2,9%)	Сущ-сущ (12%) Мест-сущ (8,9%) Сущ-глаг (6,6,%) Прил-сущ (6,6%) Сущ-предл (4,6%) Глаг-сущ (4,1%)

Статьи №1 и №2 близки по многим анализируемым показателям. Теоретически, их можно отнести к творчеству одного журналиста. Результат математического анализа в определённой степени подтверждается тем фактом, что оба текста посвящены работе Пермской городской Думы, и логично предполагать, что они написаны одним публицистом. Статья №3 посвящена американским событиям и, вполне возможно, является перепечаткой из другого российского издания, скорее всего петербургского, на что указывает выявленная лексема.

Исследование имеет огромный фронт работы: за каждый год сохранилось около 300 выпусков газеты, в каждом из которых не менее десятка статей, требующих установления авторства.

Утверждать, что результаты нашего исследования носят окончательный характер нельзя, поскольку ни один из методов атрибуции текста не даёт стопроцентной гарантии. Здесь достаточно сослаться на опыт карельских исследователей творчества Ф.М. Достоевского (Захаров В.Н., 2000), когда, гораздо более сложные тесты, приписывали великому автору статьи других людей.

В задачах оговаривалось, что установление авторства публиковавшихся материалов должно послужить базисом для составления социальных портретов журналистов. Имея портфолио из статей, мы можем воссоздать политические, социо-культурные ориентиры этих людей, попытаться соотнести полученные данные с архивными и мемуарными источниками. Таким образом, анонимные тексты могут открыть нам мир дореволюционной журналистики в рамках одной провинциальной газеты.

Источники и литература

1. Быстров А.В., Злобин Е.В. К вопросу об авторстве предсмертного письма Б.В.Савинкова - опыт комплексного исследования // Круг идей: Новое в исторической информатике. М., 1994.
2. Захаров В.Н., Рогов А.А., Сидоров Ю.В. Проблема грамматического инварианта Ф.М. Достоевского в атрибуции анонимных и псевдонимных статей журнала «Время» (1861-1863)//Труды Петрозаводского государственного университета. Сер. «Прикладная математика и информатика». Вып.9. Петрозаводск. Изд-во ПетрГу, 2000.
3. Злобин Е.В. К вопросу об авторстве "Записок" И.И.Горбачевского // История СССР. 1990. № 2.
4. Масальцева Т.Н. Литературная критика газеты «Пермские губернские ведомости» (1890-1917 годы). Автореферат дисс. [Документ MS Word]/ Пермский государственный университет. – Электрон.дан. – Пермь: Сайт Пермского государственного университета, 2006. – Режим доступа: http://www.psu.ru/psu/files/1080/Masalceva.doc.-Загл с экрана.
5. От Нестора до Фонвизина. Новые методы определения авторства/ред.Милов Л.В/. М., «Прогресс», 1994.
6. Пермские губернские ведомости. №№ 47, 49, 63. Пермь,1909.
7. Поршнева О.С., Поршнев С.В. К вопросу об атрибуции текстов записей солдатских разговоров С.З. Федорченко//Информационный бюллетень АИК. №29. М., 2002.
8. Тарнопольская. И.О. Диграммная энтропия и атрибуция анонимных текстов: результаты тестирования методики//Информационный бюллетень АИК. №23. М., 1998.