Автоматическое определение авторства

Лидия Михайловна Пивоварова

Системы понимания текста

Введение

- Определение авторства – определение одного автора из нескольких возможных
- Верификация автора – установление, принадлежит ли данный текст данному автору
- Определение плагиата – поиск сходства между двумя текстами
- Построение авторского профиля – т.е. установление пола, возраста, образования и т.п. автора конкретного текста
- Установления стилистической непоследовательности текста (что может означать, что работало несколько авторов)

Содержание

- Определение автора как задача классификации
- Методы атрибуции

Определение авторства как задача классификации

- Дано:
  - текст неизвестного автора
  - набор возможных авторов
  - примеры текстов для каждого из возможных авторов
- Задача:
  - отнести изучаемый текст к одной из представленных групп
- Вопрос в том, какие свойства использовать для классификации

Стилистические свойства

- Символьные
- Лексические
- Синтаксические
- Семантические
- Тематические

Лексические свойства текстов

- Словарный запас
  - зависит от объема текста, не может использоваться в одиночку
- Частотные распределения слов
  - текст как вектор (bag of words)
  - служебные слова (предлоги, союзы, артикли) более важны, чем значимая лексика: они используются бессознательно, их распределения сохраняются для разных тем и жанров
  - размерность пространства классификации сильно ниже, чем в тематической классификации
- N-граммы (сочетания слов)
  - не всегда улучшают качество
  - для их использования нужны большие объемы корпусов

Символьные свойства текста

- Частотные распределения букв, цифр, верхнего и нижнего регистра, знаков препинания
- N-граммы – сочетания букв
  - более устойчивы к шумам (например, опечаткам), чем лексические свойства
  - выбор N зависит от языка; чем больше N, тем больше размерность пространства классификации, тем большие нужен корпус; маленькие N (2-4) – свойства типа слогов
- Модели сжатия
  - чувствительны к тематике текстов

Синтаксические свойства

- Автор использует набор синтаксических паттернов, которые хуже осознаются, чем лексика
- Требуется синтаксический разбор текста – такой метод уже не может быть языково-независимым
- Данные всегда зашумлены (из-за несовершенства синтаксического анализа)
- Уровень анализа может быть разным:
  - Частотные распределения частей речи
  - Локальный синтаксис
  - Глобальная структура предложения
  - Словосочетания определенного типа

Семантические свойства

- Семантический анализ сам по себе менее развит, семантическая разметка дает большее число ошибок – как следствие, точность анализа снижается
- Было несколько попыток использовать семантические классы слов (WordNet) для определения авторства, однако неочевидно, что это дает преимущество по сравнению с другими методами

Тематические свойства

- Если тематика сообщений заранее известна (например, речь идет об анонимном сообщении на тематическом интернет-форуме), то можно использовать авторские предпочтения в выборе тех или иных слов, характерных для этой предметной области (доменных синонимов)
- Однако этот метод очень трудно автоматизировать – и, как следствие, переносить с одной задачи на другую

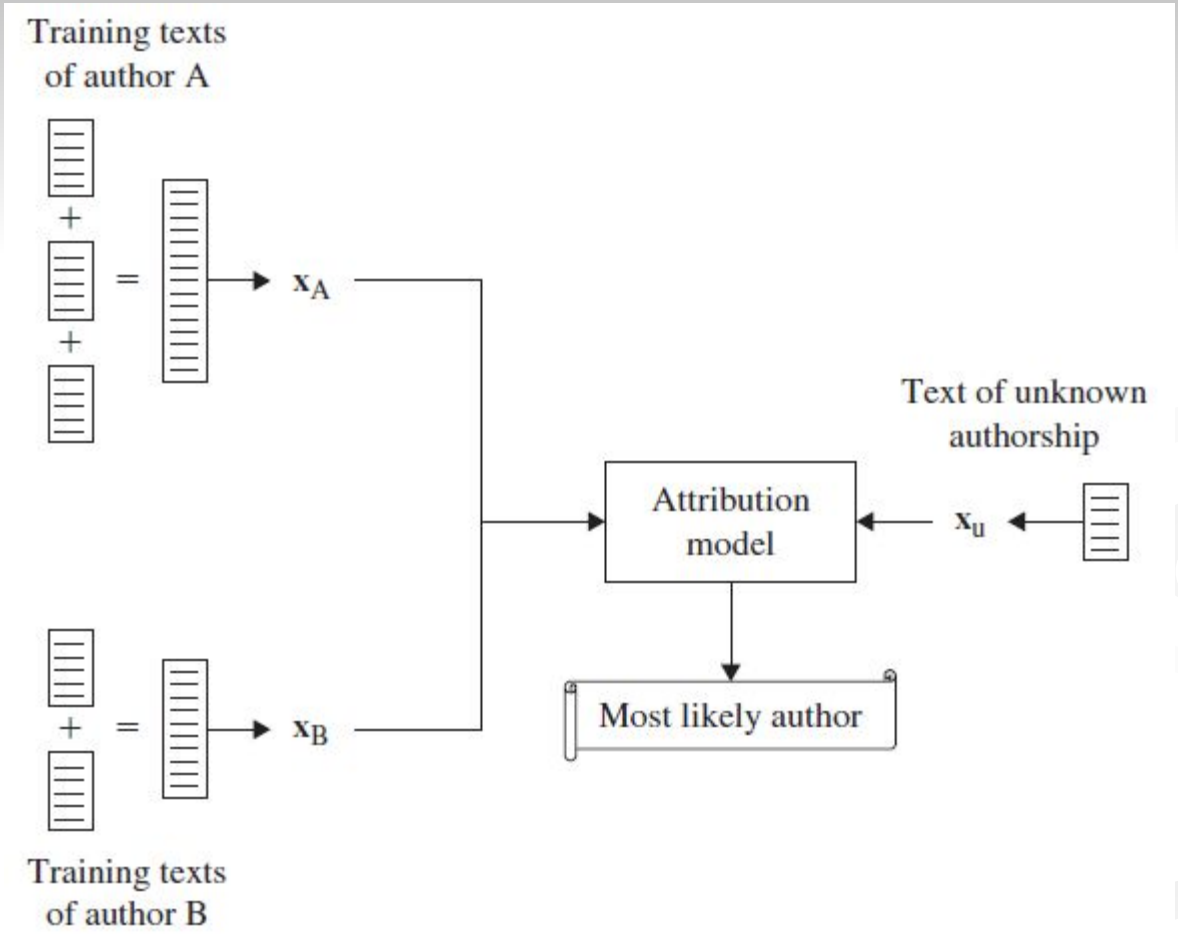
Выбор свойств

- В определении авторства лучше всего работает не одно какое-то свойство, а их сочетание
- Обычно набор свойств сначала проверяют на обучающей выборке и выбирают наиболее дискриминирующие
- Дискриминирующие свойства
  - наиболее частотны
  - наименее стабильны (т.е. имеют большое число синонимов)
- Можно использовать методы снижения размерности в пространстве слов

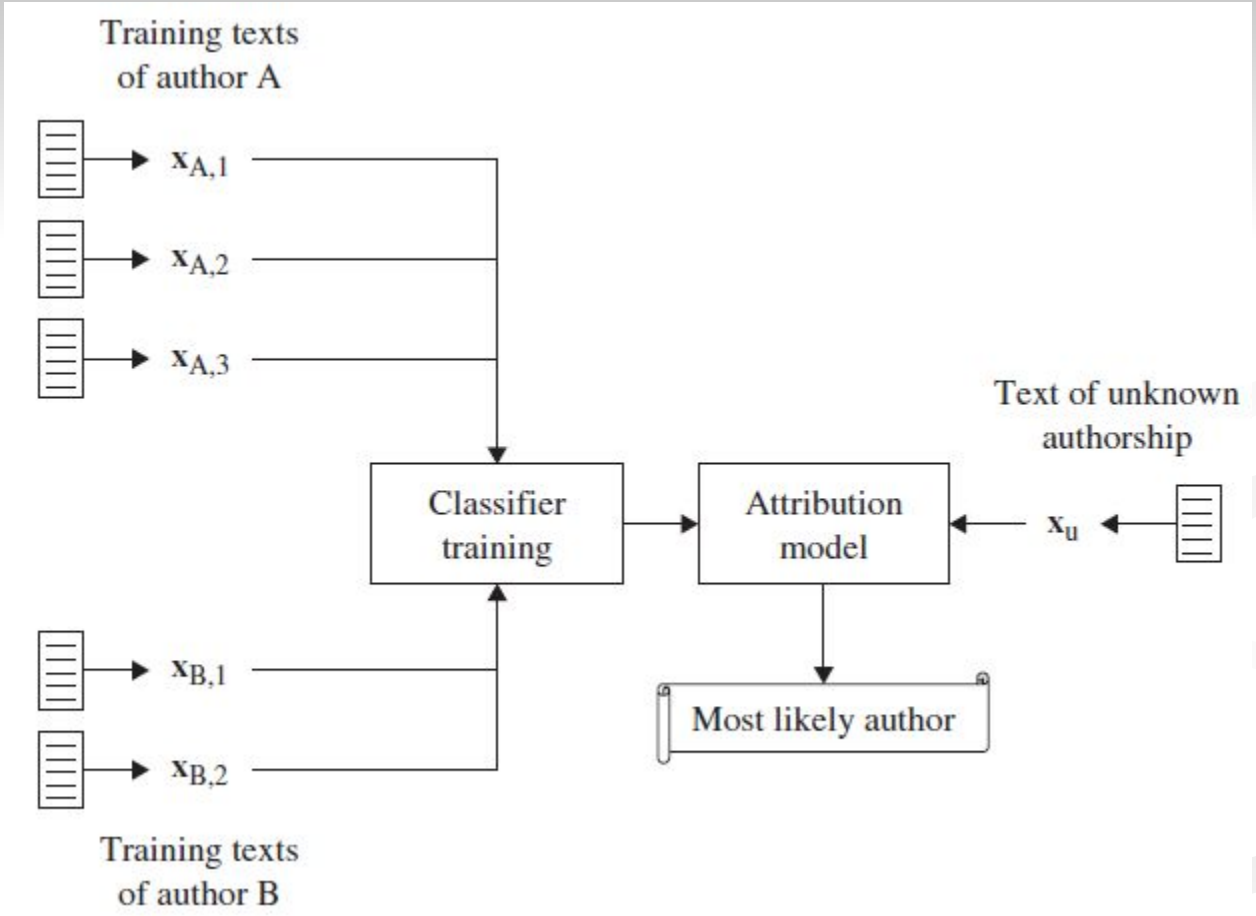
Содержание

- Определение автора как задача классификации
- Методы атрибуции

Ориентированные на автора



Ориентированные на текст



Источники

- Efstathios Stamatatos A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology Volume 60, Issue 3, pages 538–556, March 2009 - http://www.clips.ua.ac.be/stylometry/Lit/Stamatatos_survey2009.pdf