# Языконезависимое определение авторства текста на базе языковых моделей символьного уровня

## Проблема определения авторства текста

- Невыясненное авторство исторических документов
- Категоризация коллекций документов
- Извлечение информации
- Вопросно-ответные системы

#### Принцип определения авторства

Автор при составлении текста использует языковые средства различных уровней:

- Семантические
- Синтаксические
- Лексикографические
- Орфографические
- Морфологические

Особенности использованных в тексте языковых средств позволяют судить об авторстве текста.

## Возможные подходы к решению проблемы определения авторства

- Стилистический анализ
- Статистическое языковое моделирование

#### Стилистический анализ

Проводится в два этапа:

- 1) Извлечение стилевых маркеров
- 2) Построение классификатора

#### Недостатки метода

- Процедуры извлечения стилевых маркеров почти всегда зависят от языка текста
- Выбор исследуемых свойств является нетривиальной задачей
- Анализ проводится только на уровне слов
- Неприменимость к восточноазиатским языкам, в которых отсутствует явное разделение слов

## Статистическое языковое моделирование

- Заключается в выявлении закономерностей в естественном языке (семантических, лексикографических и морфологических шаблонов), на основе которых можно делать прогнозы
- Задача предсказание вероятности появления в тексте последовательностей слов, которые действительно имеют место в тексте

#### Оценка качества модели

Perplexity = 
$$\sqrt[N]{\prod_{i=1}^{N} \frac{1}{\Pr(w_i|w_1 \dots w_{i-1})}}$$

$$Entropy = log_2 Perplexity$$

#### **N-граммная модель**

Вероятность появления цепочки слов:

$$\Pr(w_1 w_2 \dots w_N) = \prod_{i=1}^{N} \Pr(w_i | w_1 \dots w_{i-1})$$

N-граммная модель аппроксимирует эту вероятность в предположении, что на вероятность появления слова влияют только последние n-1 слов:

$$\Pr(w_1 w_2 \dots w_N) = \prod_{i=1}^{N} \Pr(w_i | w_{i-n+1} \dots w_{i-1})$$

#### **N-граммная модель**

#### В самом простом случае

$$\Pr(w_i|w_{i-n+1} \dots w_{i-1}) = \frac{\#(w_{i-n+1} \dots w_i)}{\#(w_{i-n+1} \dots w_{i-1})}$$

- Использование грамм длины n означает вычисление вероятностє $W^n$  событий
- Вероятность появления новых n-грамм всегда ненулевая.

## Сглаживание вероятностных оценок

$$\begin{split} & \Pr(w_i | w_{i-n+1} \dots w_{i-1}) \\ &= \begin{cases} \widehat{\Pr}(w_i | w_{i-n+1} \dots w_{i-1}), \#(w_{i-n+1} \dots w_i) > 0 \\ \beta(w_{i-n+1} \dots w_{i-1}) \times \Pr(w_i | w_{i-n+2} \dots w_{i-1}) \end{cases} \end{split}$$

$$\widehat{\Pr}(w_i|w_{i-n+1} \dots w_{i-1}) = \frac{discount \#(w_{i-n+1} \dots w_i)}{\#(w_{i-n+1} \dots w_{i-1})}$$

#### Принципы классификации

```
Используется Баесова теория принятия
решения: текст D относится к авторской
категории c \in C = \{c_1 \dots c_{|C|}\} если
c = \arg \max_{c \in C} \{ \Pr(c|D) \}
В соответствии с правилом Байеса:
= arg max_{c \in C} \{ Pr(D|c) Pr(c) \}
= arg max_{c \in C} \{ Pr(D|c) \}
```

#### Результаты классификации

- Греческий корпус: две коллекции по 200 документов 10 различных авторов, F-мера 74% и 90%
- Английский корпус: Alex Catalogue of Electronic Texts, 8 авторов, наилучшая F-мера 98% при использовании 6-граммной модели с абсолютным сглаживанием
- 8 авторов, F-мера 94% при использовании 3граммной модели при использовании алгоритма сглаживания Виттена-Белла