ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 24 МАТЕМАТИКИ Выпуск 4
2017

А. В. Германович, С. Ю. Мельников, В. М. Хвостенко (Москва, ИСАА при МГУ, ООО «Линфо»). **О выборе множества слов, характеризующих авторский стиль арабского текста.**

Задаче определения авторства текста посвящено большое количество исследований (см. библиографический обзор в [1]). В работе [2] на материале художественных произведений XVIII–XX веков 23 русскоязычных авторов анализировалась различительная способность частот встречаемости определенных групп слов в задаче определения авторства. Группы слов подбирались лингвистами из тех соображений, что их частоты должны слабо контролироваться автором на сознательном уровне, быть устойчивыми для текстов одного автора и быть способными статистически разделить тексты разных авторов. Сделан качественный вывод, что хорошей различительной способностью для художественных прозаических текстов на русском языке обладают так называемые служебные слова (союзы, предлоги, частицы). В цитируемой работе выделено 55 служебных слов. Близкий подход, связанный с лексическим ранжированием, развивается в [3]. На сегодняшний день технологии определения авторства, основанные на методах вычислительной лингвистики, активно развиваются, в том числе для арабского языка ([4]). В настоящей работе описан подход к построению множества слов с максимальной различительной способностью, не требующий участия лингвиста.

Описание подхода. Под текстом будем понимать последовательность словоформ в алфавите языка, разделенных знаками-разделителями (пробелы, знаки препинания, неалфавитные символы). Идея подхода состоит в следующем. Пусть имеются сопоставимые по объему авторские коллекции текстов $M \geqslant 2$ авторов. По ним составляется совокупный словарь $\{w_i,\ i = 1, N\}$ встреченных в коллекции словоформ, частота которых не меньше заданного порога $K \geqslant 1$. Пусть v_{ij} — частота встречаемости словоформы $w_i \in V$ в j-й коллекции $j = 1, 2, \ldots, M$. Интуитивно понятно, что если частоты $v_{ij},\ j = 1, 2, \ldots, M,$ приблизительно равны, то словоформа обладает низкой способностью к различению авторов. Наоборот, для определения авторства важны такие словоформы, частоты которых в текстах разных авторов сильно различаются.

В качестве меры, характеризующей степень неравномерности вектора $(v_{i1}, v_{i2}, \ldots, v_{iM})$, можно выбрать статистику χ^2 :

$$\chi_i^2 = \sum_{j=1}^{M} \left(\frac{(v_{ij} - T_i/M)^2}{T_i/M} \right) \quad \text{где} \quad T_i = \sum_{j=1}^{M} v_{ij}, \quad i = 1, N.$$

Минимум этой статистики достигается на векторе с равными координатами. Словоформы (частота встречаемости которых не меньше K) с наибольшими значениями этой статистики предположительно будут наиболее информативными при различении авторов.

Результаты экспериментов с арабскими текстами. С помощью описанной в [5] системы сбора корпусов была сформирована коллекция новостных текстов

из 11 арабоязычных стран по общественно-политической и военной тематике за 2014–2017 гг. Коллекция содержит статьи 67 арабских авторов, в среднем по 12 текстов у каждого автора, средний размер каждого текста 830 словоформ. Для $K = 500$ по порогу, задаваемому для значения χ_i^2, была найдена 161 словоформа. В таблице представлены некоторые словоформы с наибольшими значениями рассматриваемой статистики.

Таблица

Словоформа	χ_i^2	Частота в коллекции	Приблизительное значение
الاحتلال	0.23	577	оккупация
فى	0.18	1291	предлог «в»
و	0.17	1906	союз «и»
ا	0.16	1078	отдельный харф
ة	0.14	586	показатель женского рода
ان	0.10	1620	служебная частица
الفلسطينية	0.09	732	палестинская
سورية	0.07	1543	Сирия / сирийская

Найденное множество словоформ разбивается на три группы: служебные слова (союзы, частицы, предлоги, местоимения) — 87 слов; слова, специфичные для политической и военной тематики (страны, движения, лидеры, военные термины) — 65 слов, другие слова — 9 слов.

Таким образом, на примере арабского языка проведенные вычисления подтверждают вывод работы [2] о хороших различительных способностях множества служебных слов в задаче идентификации автора.

СПИСОК ЛИТЕРАТУРЫ

1. *Романов А. С.*, *Шелупанов А. А.*, *Мещеряков Р. В.* Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста: Монография. Томск: В-Спектр, 2011, 188 с.
2. *Фоменко В. П.*, *Фоменко Т. Г.* Авторский инвариант русских литературных текстов. — Новая хронология Греции: Античность в средневековье. М.: Изд-во МГУ, 1996, т. 2, с. 768–820.
3. *Напреенко Г. В.* Лексико-квантитативное моделирование языковой личности в идентификационном аспекте (на материале русскоязычных интернет-дневников). Дисс. на соискание уч. степени к.ф.н. Кемерово: КГУ, 2015, 302 с.
4. *Al-Ayyoub M.*, *Alwajeeh A.*, *Hmeidi I.* An extensive study of authorship authentication of Arabic articles. — Int. J. of Web Inf. Syst., 2017, v. 13, i. 1, p. 85–104.
5. *Белозеров А. А.*, *Вахлаков Д. В.*, *Мельников С. Ю.*, *Пересыпкин В. А.*, *Сидоров Е. С.* Технологические аспекты построения системы сбора и предобработки корпусов новостных текстов для создания моделей языка. — Изв. ЮФУ. Технические науки, 2016, №12 (185), с. 29–42.