

Анализ применимости метода N-грамм в задаче определения стиля текста

12, декабрь 2015

Тарасенко Е. В.^{1,*}, Рязанова Н. Ю.¹

УДК: 81'322.2

¹Россия, МГТУ им. Н.Э. Баумана

*elenat191292@gmail.com

Введение

Одним из широко используемых методов анализа текста является метод N-грамм [1]. Он является наиболее часто применяемым в задаче обнаружения факта плагиата, так как позволяет выполнять сравнение содержания текстов [2].

Во многих задачах необходимо определить, так называемый, стиль текста. Под стилем текста понимается сложившаяся система речевых средств, используемых в различных сферах человеческого общения. В лингвистике его принято называть функциональным стилем речи [3]. Всего выделяют пять основных стилей: публицистический, официально-деловой, научный, художественный и разговорный. Стиль текста во многом определяется частотой и порядком употребления в нем различных частей речи [3], что удовлетворяет условиям применения метода N-грамм, который позволяет составлять комбинации из частей речи и определять процент их вхождения в тексты различных стилей.

1. Формирование вектора отличительных признаков

Для анализа возможности применения метода N-грамм при классификации текстов по стилям речи, необходимо сформировать вектор отличительных признаков и рассчитать его значение для каждого из стилей [4]. Такой вектор формируется из n-грамм с наиболее различающимися частотами вхождений в тексты различных стилей, т. е. их значения частот должны лежать в диапазоне значений с наибольшей длиной. Длина такого диапазона рассчитывается по формуле 1.

$$L = | \max(n_1, n_2, \dots, n_k) - \min(n_1, n_2, \dots, n_k) |, \quad (1)$$

где L – длина диапазона значений, $n_1, n_2, \dots, n_k$ – значения частот вхождений n-грамм в тексты соответствующих стилей.

Для подсчета количества n-грамм в текстах различных стилей была использована вы-

борка текстов из «Национального корпуса русского языка» (www.ruscorpora.ru). Данный корпус предоставляет набор текстов [5], классифицированный лингвистами по пяти основным стилям речи. В каждом тексте набора для каждого слова определена часть речи: существительное, прилагательное, глагол, наречие, числительное, предлог, союз, местоимение, частица или нелексический элемент [6]. Размер предоставленной выборки составил 2150 текстов, из них 427 — публицистического стиля, 430 — официально-делового, 405 — научного, 415 — художественного и 473 — разговорного.

Для каждого набора текстов одного стиля было подсчитано количество входящих в него биграмм, 3-грамм, 4-грамм и 5-грамм (n-грамм длиной два, три, четыре и пять соответственно). Из них, согласно формуле 1, были выбраны n-граммы, частота вхождений которых в тексты рассматриваемых стилей различается в наибольшей степени. Из-за очень большого числа возможных комбинаций, частоты вхождений четырех- и 5-грамм в тексты различных стилей практически совпадают, что позволяет в дальнейшем их не рассматривать.

На рисунках 1 и 2 показаны проценты вхождений биграмм и 3-грамм. Данные графики позволяют провести детальный анализ применимости метода N-грамм в задаче определения стиля текста. Анализ результатов подсчета биграмм (рисунок 1), показывает, что выделенные в текстах биграммы не могут рассматриваться в качестве отличительных признаков для текстов различных стилей, так их частота вхождений практически совпадает.

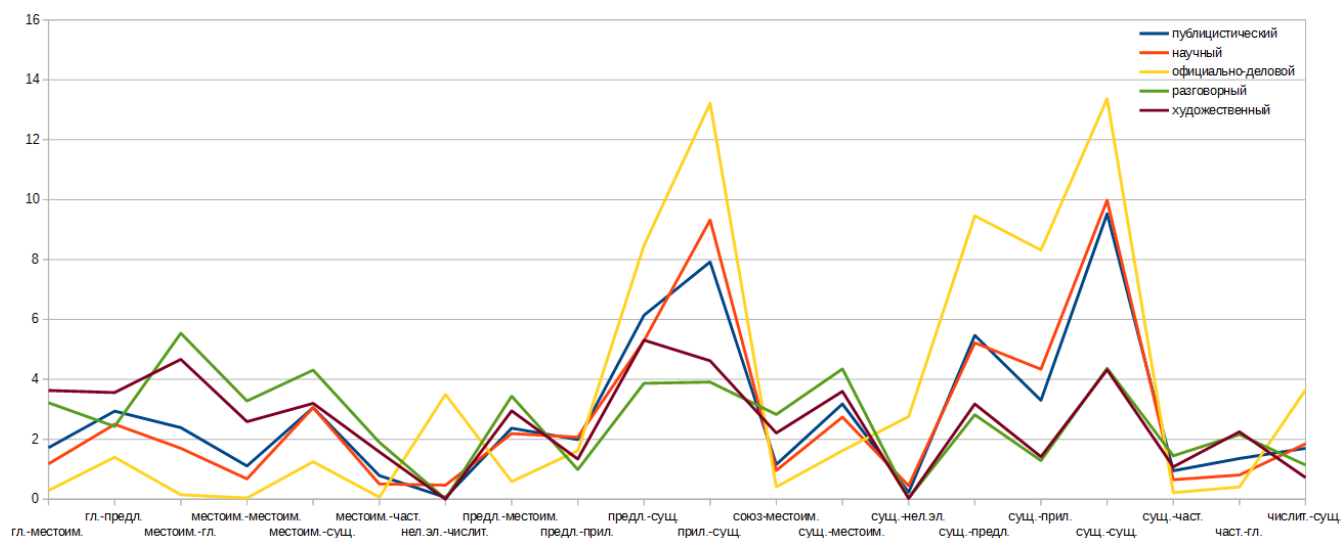


Рис. 1. График зависимости частоты вхождений биграмм от стиля текста

Анализ результатов подсчета количества 3-грамм (рисунок 2), показывает, что частоты выделенных 3-грамм совпадают по своим числовым значениям у следующих пар стилей: разговорного и художественного, публицистического и научного. Это объясняется их определенными сходствами: художественный стиль, как и разговорный, содержит большое число диалогов, монологов, эмоциональных высказываний и описаний. Публицистический стиль оперирует ограниченным числом фактов, подчиняется строгим правилам оформления, что характерно и для научного стиля. Официально-деловой, напротив, явля-

ется особым стилем, предъявляющим жесткие требования к построению оборотов речи и последовательности фраз. Таким образом, метод N-грамм можно использовать для классификации текстов по стилям речи, объединив в отдельные группы разговорный и художественный стиль и публицистический, и научный. В результате проведенных исследований были выделены следующие группы стилей:

- 1) официально-деловой;
- 2) публицистический и научный;
- 3) разговорный и художественный.

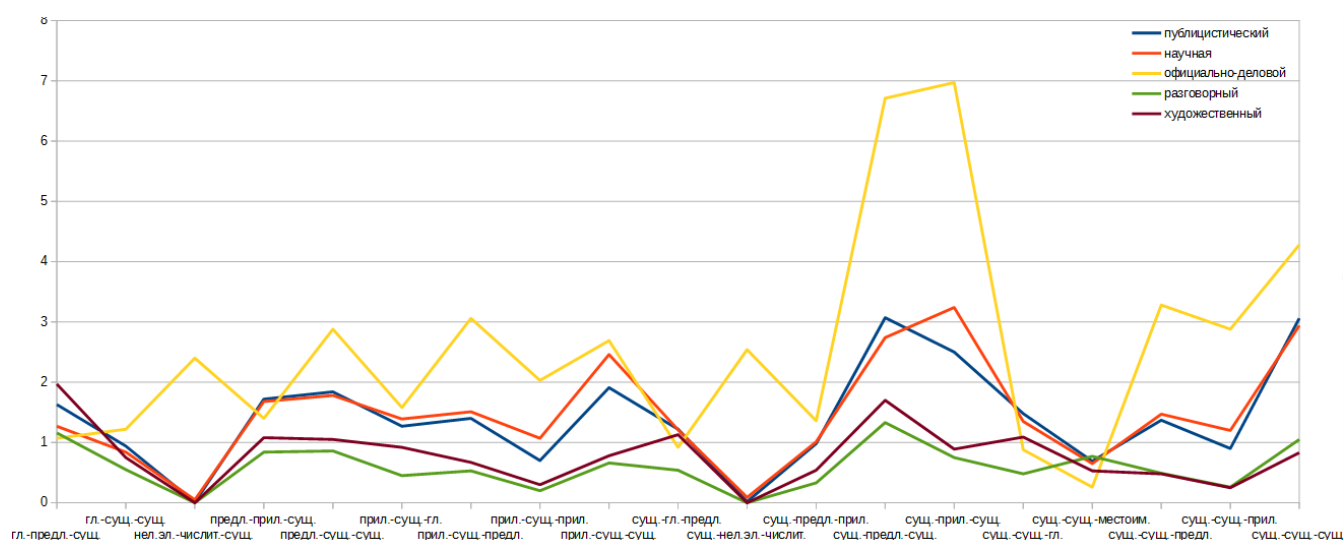


Рис. 2. График зависимости частоты вхождений 3-грамм от стиля текста

Для каждой из трех групп стилей речи были посчитаны значения вектора отличительных признаков. Полученные результаты представлены на рисунке 3. Анализ приведенных зависимостей подтверждает сделанное предположение о возможности использования метода N-грамм для классификации текстов по выделенным группам стилей.

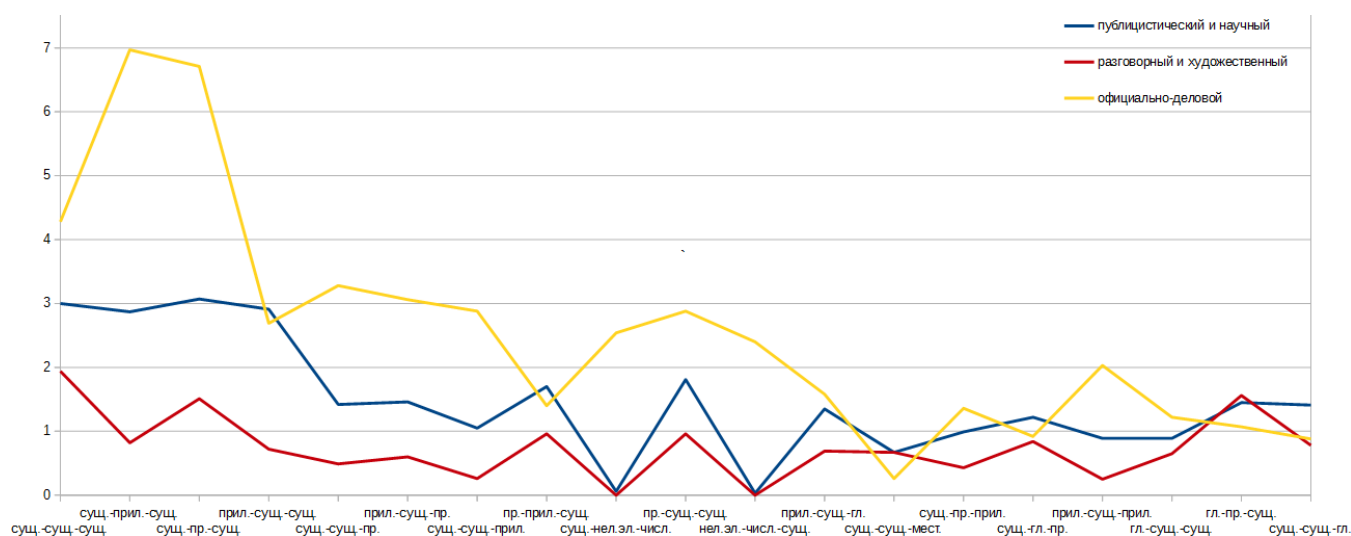


Рис. 3. График зависимости частоты вхождений 3-грамм от группы стиля

2. Определение стиля текста по полученному вектору

Для определения стиля текста необходимо подсчитать в нем количество 3-грамм, входящих в вектор отличительных признаков. Тогда стиль текста будет определяться как стиль с наиболее близкими значениями отличительных признаков. Так называемая, «близость» двух стилей вычисляется по формуле 2.

$$k = 100\% - \frac{\sum_{i=1}^n |a_i - b_i|}{n}, \quad (2)$$

где a_i – процент вхождения i -ой 3-граммы в анализируемый текст, b_i – значение соответствующего признака в сравниваемом стиле, n – размерность вектора признаков.

Для оценки правильности определения стиля речи используется точность и полнота, вычисляемые по формулам 3 и 4 соответственно [7].

$$P = \frac{|A \cap B|}{|B|} \cdot 100\%, \quad (3)$$

$$R = \frac{|A \cap B|}{|A|} \cdot 100\%, \quad (4)$$

где A – полученное множество текстов заданного стиля, B – множество текстов заданного стиля. Точность показывает процент верно классифицированных текстов заданного стиля из всех текстов данного стиля, а полнота — процент верно классифицированных текстов заданного стиля из всех текстов, которые были к нему отнесены.

Для определения точности и полноты классификации текстов по выделенным группам стилей (публицистическому и научному, официально-деловому, разговорному и художественному) была использована та же выборка текстов из «Национального корпуса русского языка». Ее повторное использование возможно так, как она не является обучающей. Полученные результаты приведены в таблице 1.

Таблица 1. Точность и полнота определения группы стиля текста

стиль речи	P , %	R , %
публицистический и научный	80	99
официально-деловой	87	70
разговорный и художественный	92	77

Полученные значения демонстрируют высокую точность классификации, которая в среднем составляет 86,3%, а также высокую полноту, составляющую в среднем 82%.

Заключение

В статье было исследовано применение метода N-грамм в задаче определения стиля текста. В результате проведенного анализа было установлено, что данный метод возмож-

но использовать для классификации текстов по трем группам стилей, объединив в одну группу публицистический и научный стили, в другую — разговорный и художественный, в третью — официально-деловой. Данный метод не позволяет отличать публицистический стиль от научного и разговорный от художественного. В результате проведенного исследования точность классификации текстов по трем выделенным группам стилей в среднем составила 86,3%, а полнота — 82%.

Список литературы

- [1]. Бузикашвили Н.Е., Самойлов Д.В., Крылова Г.А. N-граммы в лингвистике // Сборник: Методы и средства работы с документами. М.: Диториал УРРС. 2000. 376 с. С. 91-130.
- [2]. Тарануха В.Ю. Использование комбинированных критериев для автоматизированного определения заимствований // «Инновации в науке»: сборник статей по материалам XXXII международной научно-практической конференции. (Россия, г. Новосибирск, 30 апреля 2014 г.). Новосибирск: Изд. «СибАК». 2014. № 4 (29). 120 с. С. 15-18.
- [3]. Кожина М.Н., Дускаева Л.Р., Салимовский В.А. Стилистика русского языка. М.: Флинта: Наука. 2008. 464 с.
- [4]. Шевелёв О.Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений: автореф. дис. ... канд. техн. наук: 05.13.18. Томск. 2006. 18 с.
- [5]. Копотев М.В., Янда Л. Национальный корпус русского языка. (Рецензия) // Вопросы языкознания. 2006. №5. С. 149-155.
- [6]. Ляшевская О.Н., Плунгян В.А., Сичинава Д.В. О морфологическом стандарте Национального корпуса русского языка // Национальный корпус русского языка: 2003-2005. Результаты и перспективы: [сборник] / Рос. акад. наук, Ин-т рус. яз. им. В. В. Виноградова. М.: Индрик. 2005. 343 с. С. 111—135.
- [7]. Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс. 2011. 528 с.