DOI: 10.15514/ISPRAS–2021–33(1)–14

Поиск заимствований в армянских текстах путем внутреннего стилометрического анализа

Е.М. Ешилбашян, ORCID: 0000-0002-2200-7065 <yyeshilbashian@ispras.ru>
А.А. Асатрян, ORCID: 0000-0002-2529-0169 <arianasat@ispras.ru>
Ц.Г. Гукасян, ORCID: 0000-0003-2389-517X <tsggukasyan@ispras.ru>
Российско-Армянский университет,
ул. Овсепа Эмина 123, Ереван, 119991 РА

Аннотация. Работа посвящена применению внутренних стилометрических методов в задаче обнаружения текстовых заимствований для армянского языка. Мы исследуем два варианта постановки задачи: обнаружение изменения стиля в документе и обнаружение границ нарушений стиля. Для данных задач в рамках этой работы мы создаем синтетические примеры с заимствованиями для академического, художественного и новостного жанров текста, и на полученных примерах проверяем эффективность алгоритмов иерархической кластеризации и других моделей по обнаружению нарушений стиля из серии конференций PAN.

Ключевые слова: стилометрический анализ; обнаружение текстовых заимствований

Для цитирования: Ешилбашян Е.М., Асатрян А.А., Гукасян Ц.Г. Поиск заимствований в армянских текстах путем внутреннего стилометрического анализа. Труды ИСП РАН, том 33, вып. 1, 2021 г., стр. 209-224. DOI: 10.15514/ISPRAS–2021–33(1)–14

Благодарности. Авторы благодарят Недумова Я.Р., Скорнякова К.А. и Турдакова Д.Ю. за ценные отзывы и обсуждения.

Plagiarism Detection in Armenian Texts Using Intrinsic Stylometric Analysis

Ye.M. Yeshilbashian, ORCID: 0000-0002-2200-7065 <yyeshilbashian@ispras.ru>
A.A. Asatryan, ORCID: 0000-0002-2529-0169 <arianasat@ispras.ru>
Ts.G. Ghukasyan, ORCID: 0000-0003-2389-517X <tsggukasyan@ispras.ru>
Russian-Armenian University,
123 Hovsep Emin str., Yerevan, 0051 Armenia

Abstract. In this work we study the application of intrinsic stylometric methods to the task of plagiarism detection in Armenian texts. We use two task setups from PAN's series of conferences on text forensics and stylometry: style change detection and style breach detection. Style change detection aims to determine whether the text is written by more than one author, while style breach detection detects the boundaries of stylistically distinct text fragments. For these tasks, we generate synthetic test sets for three genres of text: academic, literature, and news, and then use them to evaluate the effectiveness of hierarchical clustering and other relevant models from PAN conferences. We employ a standard set of character-level, lexical and readability features, and additionally perform morphological and dependency parsing of text fragments to extract syntactic features encoding author style information. The evaluation results show that the clustering-based approach fails to correctly detect style change detection in longer texts and is only marginally better for shorter texts. For style breach detection, hierarchical clustering-based approach performs better than a random baseline classifier, but the difference is not sufficient to warrant its practical use. In a complementary experiment, we show that reducing the number of features and multicollinearity in them via PCA helps to increase the precision of style breach detection methods for certain text categories.

Keywords: stylometric analysis; plagiarism detection

For citation: Yeshilbashian Ye.M., Asatryan A.A., Ghukasyan Ts.G. Plagiarism Detection in Armenian Texts Using Intrinsic Stylometric Analysis. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 1, 2021, pp. 209-224 (in Russian). DOI: 10.15514/ISPRAS–2021–33(1)–14

Acknowledgements. The authors thank Ya.R. Nedumov, K.A. Skorniakov and D.Yu. Turdakov for insightful feedback and discussions.

1. Введение

Методы поиска заимствований разбиваются на внешние и внутренние. Внешние методы сравнивают текст проверяемого документа с проверочным набором потенциальных источников. При отсутствии проверочной базы применяются внутренние методы поиска заимствований, которые основываются исключительно на имеющемся документе для нахождения подозрительных участков в нем. Так как для армянского языка наличие такой базы ограничено, и многие заимствования, встречаемые на практике, являются переводом текстов на других языках, то становится актуальным исследование возможности применения внутренних методов поиска заимствований.

Для установления наличия заимствований в тексте, методы внутреннего анализа подвергают его стилометрическому анализу. В этом контексте стиль текста определяется закономерностью использования словоформ, знаков препинания, стиля речи, а также уровнем читабельности. Предполагается, что каждый автор обладает собственным уникальным стилем написания текста, и фрагменты, где наблюдается отхождение от этого стиля, предположительно являются заимствованиями из работ других авторов.

В рамках данной работы мы проверяем на армяноязычных текстах применимость существующих моделей обнаружения нарушений стиля в документах. Эффективность моделей оценивается на автоматически сгенерированных наборах примеров из трех жанров текстов – академических, художественных, и новостных. Мы представляем результаты для двух конфигураций постановки задачи: бинарная классификация, отвечающая на вопрос, содержит ли текст нарушения стиля, и определение границ стилистически однородных участков в тексте. Наборы данных, исходные код и другие ресурсы, необходимые для тестирования моделей, доступны на GitHub[1].

2. Обзор литературы

Исследованию и разработке стилометрических методов были посвящены несколько конференций PAN. В них встречаются разные постановки задач внутреннего анализа, из которых основными являются обнаружение изменения стиля в документе (style change detection) [1-2], обнаружение границ нарушений стиля (style breach detection) [3], кластеризация по авторству (author clustering) [3-4].

2.1 Обнаружение изменения стиля

Обнаружение изменения стиля формулируется как задача классификации, где нужно определить содержит ли текст отклонения от преобладающего в нем стиля. Для обзора существующих методов по обнаружению изменения стиля текста были изучены работы участников конференции PAN 2018 и 2019 годов [1-2]. На тестовых выборках наилучшие результаты показали модели Натха (Sukanya Nath) [5], Златковой (Dimitrina Zlatkova) и др.

[1] https://github.com/ivannikov-lab/style-change-analysis

[6], Хосснйнии (Marjan Hosseinia) и Мукерджи (Arjun Mukherjee) [7], а также Сафина (Kamil Safin) и Огальцова (Aleksandr Ogaltsov) [8].

В [6] документ делится на несколько сегментов, для каждого из которых выполняется извлечение лексических, синтаксических признаков, и далее на каждом из этих групп признаков применяются 4 классификатора – SVM, случайный лес, AdaBoost и многослойный персептрон. Каждому классификатору присваивается вес на основе достоверности предсказания, и на основе этого далее каждой группе признаков сопоставляется вектор с вероятностями предсказания класса. Эти векторы вместе с результатами градиентного бустинга на основе TF-IDF подаются на вход логистической регрессии, результатом которой является ответ, присутствует ли в тексте изменение стиля или нет.

Авторы [5] использует метод, основанный на алгоритме пороговой кластеризации. Для извлечения признаков документ делится на окна, и каждое окно преобразуется в вектор признаков на основе нормализованной частоты выбранных токенов. Для определения расстояния между векторами авторы выбрали меру расстояния Матусита. Идея алгоритма пороговой кластеризации состоит в том, чтобы построить список расстояний между окнами и итеративно выбирать окна с наименьшим расстоянием так, чтобы при формировании кластера ближайшие элементы включались первыми. После этого включаются окна, которые находятся дальше. Результатом алгоритма является число кластеров, которое обозначает количество авторов в проверяемом тексте.

Метод [7] основан на идее определения изменения стиля в тексте на основе синтаксического анализа. Для каждого предложения в документе строится синтаксического дерево разбора, далее последовательность этих деревьев передается на вход двух рекуррентных нейронных сетей, первая из которых обрабатывает последовательность в прямом порядке, как в документе, а вторая – в обратном. Далее, при помощи функции схожести оценивается разность между результатами сетей, на основе которой определяется вероятность изменения стиля текста. Так как данная модель сильно зависима от качества синтаксического анализа текстов, а в случае армянского языка точность таких анализаторов сравнительно низкая, она не рассматривалась в рамках данной работы.

В работе [8], как и в модели [6], используется ансамбль классификаторов, каждый из которых по некоторому признаку определяет, встречается ли в документе изменение стиля текста или нет. Первый классификатор использует 19 статистических признаков (число предложений, частота уникальных слов, длина текста и т.д.), для второго классификатора каждый текст представляется в виде 3000-мерного вектора, который содержит информацию о частоте символьных n-грамм в тексте. Третий классификатор применяется на векторном представлении текста, который содержит векторные представления n-грамм (n=1, …, 6), и размерность которого составляет более 3 миллионов. В конце вычисляется взвешенная сумма результатов классификаторов, которая используется как оценка вероятности присутствия в тексте изменения стиля.

2.2 Обнаружение границ нарушений стиля

Задача выявления границ нарушений стиля заключается в определении моно- или мульти-авторства исследуемого документа и разбиении документа на стилистически однородные фрагменты в случае мульти-авторства с указанием границ смен стиля. Данная задача была предложена участникам PAN в 2017 году [3]. Лучший результат показал метод Карася (Daniel Karaś) и др. [9], в котором документ разбивается по параграфам, а затем каждый параграф представляется в виде вектора признаков, полученного путем конкатенации tf-idf значений по униграммам слов, по 3-граммам, по стоп-словам, по размеченным частям речи и пунктуационным символам. Для выявления отклонений от стиля используется статистический подход – тест Вилкоксона над попарными векторами-строками. Наименьшие

30% значений тестов считаются аномальными и ставится граница в начало крайнего параграфа соответствующего теста.

Другие методы из этой конференции, Хана (Jamal Ahmad Khan) и др. [10], а также Сафина и др. [11] были менее эффективны в данной задаче, и поэтому не рассматривались в нашей работе для адаптации и применения к армянскому языку. В этих методах сравниваются представления рядом стоящих предложений. В [10] документ разделяется по предложениям, и рассматриваются скользящие окна из трех предложений, имеющие одно общее предложение. Они используют статистические признаки на основе синтаксических, лексических характеристик текста. Алгоритм [11] применяет модель Skip-Thought, основанную на предсказании предыдущего и следующего предложения, с архитектурой кодировщик-декодировщик и с использованием GRU сетей. Предложение считается отклонением, если среднее расстояние его векторного представления от остальных векторов больше заданного допустимого. Помимо того, что данная модель работала значительно медленно по сравнению с остальными, ее адаптация к армянскому языку была бы проблематичной также потому, что для предобучения и эффективной работы используемой нейронной сети понадобилось бы большое количество текстовых данных.

2.3 Кластеризация по авторству

Задача авторской кластеризации заключается в группировании небольших моно-авторских документов по группам авторов. В рамках данной работы мы рассматривали модели, предназначенные для решения той вариации этой задачи, в которой нужно каждый документ необходимо присвоить кластеру одного автора. Данная задача авторской кластеризации рассматривалась в рамках открытых соревнований PAN 2016 и 2017 [3-4].

В условиях неизвестности точного количества авторов лучший результат показали алгоритмы Гомес-Адорно (Helena Gómez-Adorno) и др. [12], Гарсии-Мондеха (Yasmany García-Mondeja) и др. [13], а также Кохера (Mirco Kocher) и Савоя (Jacques Savoy) [14], использующие иерархическую кластеризацию. В модели [12] для представления документа используются n-граммы слов и символов, над которыми далее применяется иерархическая агломеративная кластеризация, где количество кластеров определяется путем максимизации индекса Калинского-Харабаша (Calinski-Harabaz score, отношение среднего значения дисперсии между кластерами и дисперсия внутри кластера).

В [13] для векторного представления документов используется модель мешка слов. Схожести между векторами определяется по трем функциям расстояния – косинусное, Дайса и Жаккара. Они использовали β-compact графовую кластеризацию – метод построения ориентированного графа, где вершины i и j соединяются в один кластер, если они являются β -схожими и вершина j наиболее схожая из всех остальных вершин. Порог схожести β определяется на стадии тренировки, при котором максимизируется метрика FBcubed.

В своем решении авторы [14] используют иерархическую кластеризацию Single-link, где для слияния текущих кластеров сравниваются их наиболее близкие вектора. Для представления документа вектором алгоритм использует частоты наиболее частых токенов (слов и пунктуации) и наиболее частых символьных 6-грамм. Для обозначения схожести используется Канберровское расстояние, и для сравнения полученных значений вводится понятие «достаточно маленькое расстояние».

Все указанные алгоритмы используют некую парадигму «снизу-вверх», оценивая схожесть всех пар документов и используя эти значения для последовательного формирования кластеров. Применение методов кластеризации для группировки небольших текстов по стилю также было исследовано в работе [15], где применяется метод k-средних на коллекции анонимных электронных писем. С учетом того, что документы, на которых проверялись эти модели кластеризации по авторству, имели размер, сравнимый с параграфами текста, мы

посчитали целесообразным применение этих моделей в задаче определения нарушений границ стиля.

3. Методы

Для определения изменения стиля в данной работе мы рассматриваем алгоритм [5], а для задачи нахождения границ нарушений стиля – алгоритм [9] и иерархическую модель кластеризации. В рассматриваемых моделях мы в качестве единиц стилистически однородных участков текста используем параграфы, считая маловероятной нарушения стиля внутри одного параграфа [16]. Соответственно, в алгоритме [9] и иерархической кластеризации документ представляется как список параграфов.

Для кластеризации применяется агломеративная модель, в которой новые кластеры создаются рекурсивным объединением схожих мелких кластеров. Схожесть между кластерами определяется по методу Варда как прирост суммы квадратов расстояний объектов до центра кластера в случае их объединения. На каждом шаге алгоритма объединяются те два кластера, которые приводят к минимальному увеличению дисперсии. В результате работы алгоритма каждому параграфу присваивается номер кластера. Для обнаружения нарушения стиля между параграфами последовательно сравниваются номера их кластеров. В случае если два соседних параграфа принадлежат разным кластерам, между ними ставится граница смены стиля.

2.1 Признаки

В задаче определения изменения стиля модель [5] используется в своем оригинальном варианте, без изменений набора признаков. В задаче нахождения границ нарушений стиля в модели [9] и кластеризации, помимо общепринятых распространенных признаков из решений PAN, мы дополнительно извлекаем собственный набор признаков для описания синтаксических особенностей параграфа. Эффективность синтаксической информации в стилометрических задачах была установлена в работах [17-19]. В целом, используются различные символьные, лексические, морфологические и синтаксические признаки (полный список приводится в Приложении А).

1. Символьные признаки: мы извлекаем две группы символьных признаков: первая группа описывает наличие и частоту используемых суффиксов и префиксов, вторая - описывает наличие и частоту знаков пунктуации, а также стиль их комбинирования с пробелами (например, наличие пробела перед знаком пунктуации).
2. Лексические признаки:
   - общие характеристики, такие как наличие неформальных слов, терминов на латинском/кириллице, наличие и частота использования длинных и коротких слов;
   - использование сокращений, использование сокращений вместо полных форм, предпочтительные формы сокращений и стиль их склонения, а также выбор заглавных или строчных букв при написании;
   - формы написания количественных и порядковых числительных, стиль написания дат;
   - наличие и частота нетипичных для других документов n-грамм слов (с низким значением idf, вычисленном на коллекции текстов [20]).
3. Синтаксические признаки:
   - общие закономерности на основе длин предложений (отдельно рассматривается и посимвольная длина, и количество слов в предложении), как, например, минимальная, средняя, максимальная длина предложений, наличие и частота длинных и коротких предложений, количество предложений в параграфе;
   - предпочтения к использованию определенных частей речи, наличие и частота использования конкретных 2-грамм и 3-грамм частей речи;

   - наличие и частота конкретных синтаксических зависимостей в предложении, использование простых или сложных предложений;
4. Читабельность: индексы читабельности текста, адаптированные к армянскому языку: Flesch reading ease, Flesch-Kincaid, SMOG, Coleman-Liau, automated readability index, Dale-Chall, difficult words, Linsear write formula и Gunning fog.

Для токенизации текста на слова и предложения использовалась библиотека UDPipe 2.0[2] Для морфологического и синтаксического анализа использовалась библиотека Stanza[3].

3. Эксперименты

Мы провели эксперименты, чтобы оценить качество работы алгоритмов определения изменения стиля и границ нарушений стиля в документе. Для этих экспериментов мы создали автоматически сгенерированные синтетические тестовые примеры для трех жанров текстов: академического, художественного и новостного. Также, чтобы оценить эффективность этих алгоритмов, их результаты на этих данных были сравнены со случайными и тривиальными базовыми моделями.

3.1 Наборы данных

Для построения набора данных были использованы документы с одинаковой тематикой, написанные разными авторами. Построение каждого примера происходило автоматически, путем объединения параграфов исходных документов. Один из документов выбирался в качестве основного, и далее некоторые из его параграфов случайным образом заменялись на параграфы других документов одинакового содержания. В процессе генерации примеров вероятность замены параграфа на заимствование из другого документа контролировалась с целью, чтобы получить примеры с разным содержанием заимствований. Также многие сгенерированные примеры в итоге были отфильтрованы, чтобы избежать присутствия очень близких примеров в полученном наборе.

В качестве исходных данных выбирались следующие наборы данных.

1. Защищенные кандидатские диссертации РА.

   Работы, представленные в открытом доступе на официальном сайте[4], доступны только в формате PDF. Мы обработали извлеченные тексты (удалили вводные главы, списки литературы, сокращений, нумерацию, сноски, номера страниц) и далее сгруппировали на основе направления работы. При выборе данного корпуса мы основывались на предположении, что все работы написаны одним автором.
2. Энциклопедические статьи.

   В качестве источников для генерации примеров мы использовали статьи об известных личностях из Википедии и онлайн-версии Армянской энциклопедии[5].
3. Учебники.

   В качестве другого источника примеров академического жанра, были выбраны материалы из учебники старших школ и университетские пособия. Были выбраны учебники по истории армянского народа 7-9 классов на армянском языке и пособие «История Армении» для вузов[6].
4. Художественные тексты

   Для генерации примеров с текстом из литературного жанра мы использовали

[2] http://ufal.mff.cuni.cz/udpipe/2
[3] https://stanfordnlp.github.io/stanza/
[4] http://etd.asj-oa.am/
[5] http://www.encyclopedia.am/
[6] https://lib.armedu.am/

произведения, которые имеют одинаковую сюжетную линию, но написаны разными авторами: произведения «Фазан» Акселя Бакунца и сценарий к фильму «Этот зеленый, красный мир», написанный Грантом Матевосяном, по мотивам данного произведения;

5. Новостные статьи

   Чтобы найти новостные тексты, описывающие одно и то же событие, но из разных источников, мы использовали агрегатор новостей NewsHub[7].

В результате было сгенерировано 144 примера из учебников, 84 новостных, 50 художественных, 400 из диссертаций и 44 из энциклопедий (табл. 1).

Табл. 1. Количество примеров в сгенерированных тестовых наборах
Table 1. The number of examples in the generated test cases

Жанр		Количество примеров								
		Все	Количество параграфов				Процент заимствований			
			5	10	25	50	0	0-20	20-40	40-50
Академический	Диссертации	400	100	100	100	100	70	82	173	75
	Учебники	144	14	33	46	51	10	41	65	22
Художественный		50	11	22	17		10	14	19	6

3.2 Результаты

Обнаружение изменения стиля: в данной конфигурации мы протестировали алгоритмы [5], [6] и, дополнительно, алгоритмы обнаружения границ нарушений стиля – [9] и кластеризацию. Последние три модели почти на всех примерах давали положительный ответ, то есть фактически работали как тривиальный классификатор. По этой причине в данном разделе более подробно изучаются только результаты модели [5] (Таблица 2). Для каждого тестового набора результаты приводятся отдельно. Учитывая преобладание положительных примеров в наборах, для получения полной картины поведения алгоритмы также изучается доля ложно положительных предсказаний, помимо точности и полноты.

Табл. 2. Качество обнаружения изменения стиля модели [5]
Tab. 2. The quality of the model [5] in detection of changes in the style

Жанр текстов		Точность	Полнота	F1	Специфичность	Доля ложно положительных (FPR)
Академический	Диссертации	0.9002	0.8105	0.853	0.3432	0.6567
	Учебники	0.9217	0.8217	0.8688	0.4	0.6
	Энциклопедии	0.4074	0.55	0.468	0.3333	0.6666
Художественный		0.8437	0.675	0.75	0.5	0.5
Новости		0.0417	0.1428	0.0645	0.7012	0.2987

Среди академических текстов алгоритм работает с примерно одинаковым уровнем доли ложно положительных, и дает правильные предсказания лишь на около трети полностью оригинальных примеров. В этом плане заметно лучше результаты для новостных текстов, доля ложно положительных ниже 30%. Однако для этих примеров присутствует другая проблема: алгоритм показывает склонность к классификации большинства новостных примеров как полностью оригинальных. Для энциклопедических примеров также присутствует данная проблема.

[7] https://newshub.am/

Чтобы понять насколько действительно эффективен алгоритм, мы сравниваем его со случайным классикатором (рис. 1). Для сравнения качестве метрики используется специфичность, которая показывает какой процент действительно оригинальных документов был предсказан таковым алгоритмом. Данный анализ показал, что для длинных документов (с количеством параграфов 25 и более), независимо от жанра текста результаты алгоритма работает значительно хуже базового метода. Это связано с тем, что с ростом размера текста алгоритм показывает тенденцию находить присутствие заимствований.

Дополнительно мы также изучили влияние количества заимствований в тексте документа на качество предсказаний (рис. 2). Для всех категорий примеров наблюдается четкая корреляция роста качества предсказания с ростом процента заимствований.

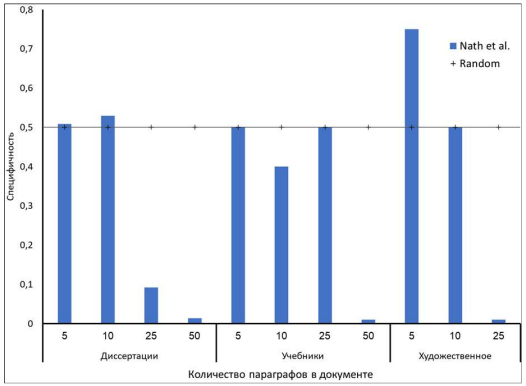

Рис. 1. Уровень специфичности обнаружения изменения стиля в тексте в зависимости от его жанра и длины
Fig. 1. The level of specificity of detecting changes in style in the text depending on its genre and length

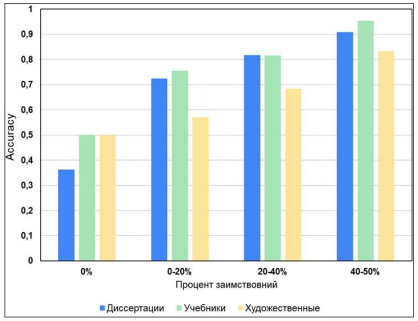

Рис. 2. Зависимость accuracy от процента заимствований для модели [5]
Figure: 2. Dependence of accuracy on the percentage of borrowing for the model [5]

Обнаружение границ нарушений стиля: в данной конфигурации мы протестировали алгоритмы [9] и агломеративную кластеризацию (AC). Аналогично [2], мы тоже в качестве базового метода используем случайный классификатор, однако с немного другой конфигурацией: в наших экспериментах мы используем базовый метод, который для каждой границы параграфов с вероятностью 0.25 предсказывает присутствие нарушения стиля.

Для оценки качества этих моделей используются точность (WinP) и полнота (WinR), заданные по формулам:

$$WinP = \frac{TP}{TP + FP} \quad (1)$$

$$WinR = \frac{TP}{TP + FN} \quad (2)$$

Точность определяется как соотношение правильно поставленных моделью границ от всех границ (TP), поставленных моделью (TP+FP), а полнота – соотношение правильно поставленных моделью границ от всех реальных границ (TP+FN).

Результаты алгоритмов для каждого тестового набора иллюстрированы на рис. 3. Чтобы определить, насколько уверенно можно утверждать превосходство [9] и AC над случайным классификатором, мы вычислили 90% доверительный интервал. В целом, почти на всех текстах, кроме художественных, кластеризация обошла [9]. Также, для художественных текстов преимущество данного алгоритма над базовым методом несущественное, а для энциклопедий и новостных текстов можно говорить, что его качество заметно хуже. Кроме диссертаций и художественных текстов, алгоритм [9] показал результаты хуже базового метода. В случае учебников и энциклопедий его точность оказалась близкой к нулю. Для этих категорий текстов данный алгоритм редко находил границы нарушений стиля.

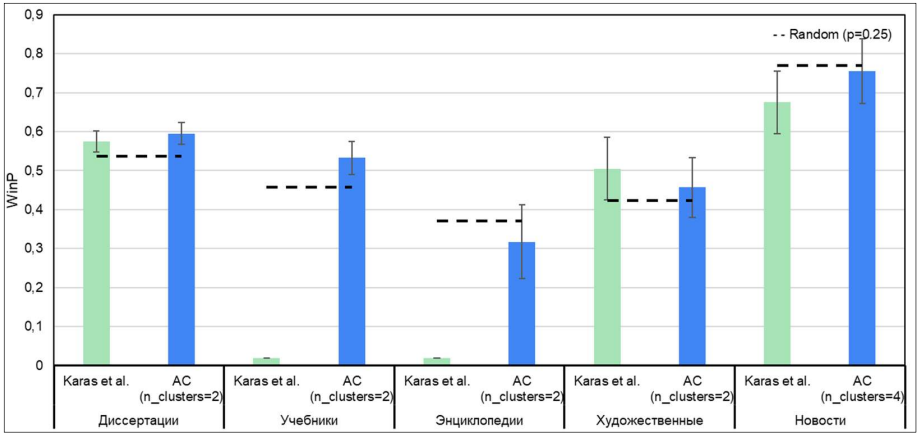

Рис. 3. Сравнение наиболее точных (90% доверительный интервал) моделей обнаружения границ нарушений стиля и случайного классификатора для каждого жанра
Fig. 3. Comparison of the most accurate (90% confidence interval) models for detecting the boundaries of style violations and a random classifier for each genre

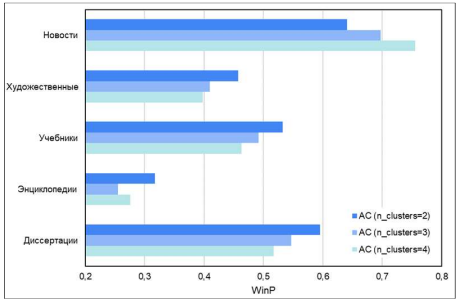

Рис. 4. Зависимость точности модели AC от количества кластеров
Fig. 4. Dependence of the accuracy of the AC model on the number of clusters

Для алгоритма кластеризации мы проверяли качество для количества кластеров 2, 3, и 4 (рис. 4). За исключением новостных текстов, в среднем самая высокая точность была получена при

ограничении количества кластеров на 2. С учетом наличия многих тесно связанных признаков в описаниях параграфов, а также ограниченного количества параграфов в отдельном документе, для повышения точности работы мы также протестировали сокращение размерности признакового описания с помощью PCA, следуя примеру [21], и для художественных текстов, учебников и диссертаций получили заметные улучшения в точности (рис. 5).

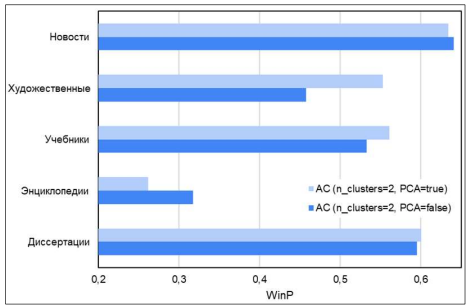

Рис. 5. Влияние PCA на точность модели AC
Figure: 5. Influence of PCA on the accuracy of the AC model

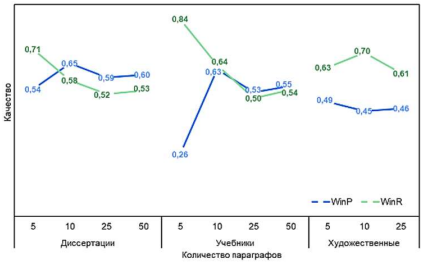

Рис. 6. Зависимость качества обнаружения границ изменений стиля от длины документов для модели AC (n_clusters=2)
Fig. 6. Dependence of the quality of detection of boundaries of style changes on the length of documents for the AC model (n_clusters = 2)

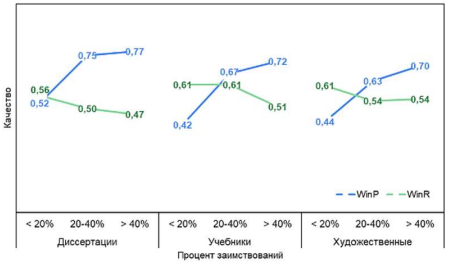

Рис. 7. Зависимость качества обнаружения границ нарушений стиля от процента заимствований в документах для модели AC (n_clusters=2)
Fig. 7. Dependence of the quality of detection of style violation boundaries on the percentage of borrowings in documents for the AC model (n_clusters = 2)

Дополнительно было изучено влияние длины текста и количества заимствований на качество предсказаний алгоритма кластеризации (рис. 6 и рис. 7). С точки зрения полноты результатов, наблюдается закономерность по его уменьшению параллельно с ростом длины документа и процента заимствований. В случае точности нет явной корреляции между значениями по этой

метрике и количеством параграфов в тексте. Только в случае текстов учебников наблюдается аномально низкая точность для маленьких документов с 5 параграфами. При этом чем меньше заимствованных участков текста в документе, тем менее точно работает алгоритм.

4. Заключение

В данной работе была исследована эффективность стилометрических методов внутреннего анализа для нахождения заимствований в текстах на армянском языке. Были изучены для трех категорий текстов (академических, художественных, новостных) и протестированы алгоритмы, основанные на лучших моделях с соревнований PAN 2017-2019. Было установлено, что для длинных документов алгоритмы нахождения изменения стиля на основе кластеризации показывают низкую специфичность и поэтому неэффективны. В задаче определения границ нарушений стиля для конкретных категорий текстов алгоритмы достигают качества выше случайного классификатора. Кроме того, была установлена эффективность применения PCA на входных признаках для сокращения их размерности.

В качестве направления дальнейших исследований рассматривается поиск более эффективных и точных наборов признаков. Дополнительным обоснованием необходимости в таком исследовании служит тот факт, что текущие синтаксические анализаторы армянского языка все еще сравнительно отстают от state-of-the-art для аналогичных языков.

Список литературы / References

[1]. Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 2125, 2018.

[2]. Eva Zangerle, Michael Tschuggnall, Günther Specht, Martin Potthast, and Benno Stein. Overview of the Style Change Detection Task at PAN 2019. Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 2380, 2019.

[3]. Michael Tschuggnall, Efstathios Stamatatos, Ben Verhoeven, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. Overview of the Author Identification Task at PAN 2017: Style Breach Detection and Author Clustering. Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 1866, 2017.

[4]. Paolo Rosso, Francisco Rangel, Martin Potthast, Efstathios Stamatatos, Michael Tschuggnall, and Benno Stein. Overview of PAN 2016 – New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. Lecture Notes in Computer Science, vol. 9822, 2016, pp. 332-350.

[5]. Sukanya Nath. Style Change Detection by Threshold Based and Window Merge Clustering Methods. Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 2380, 2019.

[6]. Dimitrina Zlatkova, Daniel Kopev, Kristiyan Mitov, Atanas Atanasov, Momchil Hardalov, Ivan Koychev, and Preslav Nakov. An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection – Notebook for PAN at CLEF 2018. Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 2125, 2018.

[7]. Marjan Hosseinia and Arjun Mukherjee. A Parallel Hierarchical Attention Network for Style Change Detection – Notebook for PAN at CLEF 2018. Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, vol. 2125, 2018.

[8]. Kamil Safin and Aleksandr Ogaltsov. Detecting a Change of Style Using Text Statistics – Notebook for PAN at CLEF 2018. Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 2125, 2018.

[9]. Daniel Karaś, Martyna Śpiewak, and Piotr Sobecki. OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection – Notebook for PAN at CLEF 2017. Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum, vol. 1866, 2017.

[10]. Jamal Ahmad Khan. Style Breach Detection: An Unsupervised Detection Model – Notebook for PAN at CLEF 2017. Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 1866, 2017.

[11]. Kamil Safin and Rita Kuznetsova. Style Breach Detection with Neural Sentence Embeddings – Notebook for PAN at CLEF 2017. Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 1866, 2017.

[12]. Helena Gómez-Adorno, Yuridiana Alemán, Darnes Vilariño Ayala, Miguel A. Sanchez-Perez, David Pinto, and Grigori Sidorov. Author Clustering using Hierarchical Clustering Analysis – Notebook for PAN at CLEF 2017, Working Notes of CLEF 2017 м Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 1866, 2017.

[13]. Yasmany García-Mondeja, Daniel Castro-Castro, Vania Lavielle-Castro, and Rafael Muñoz. Discovering Author Groups using a B-compact graph-based Clustering – Notebook for PAN at CLEF 2017. Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 1866, 2017.

[14]. Mirco Kocher and Jacques Savoy. UniNE at CLEF 2017: Author Clustering – Notebook for PAN at CLEF 2017. Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 1866, 2017.

[15]. Iqbal Farkhund, Hamad Binsalleeh, Benjamin C.M. Fung, and Mourad Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. Digital Investigation, vol. 7, issue 1-2, 2010, pp. 56-64.

[16]. Zuo Chaoyuan, Yu Zhao, and Ritwik Banerjee. Style Change Detection with Feed-forward Neural Networks. Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR Workshop Proceedings, vol. 2125, 2019.

[17]. Hirst Graeme, and Ol'ga Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing, vol. 22, no. 4, 2007, pp. 405-417.

[18]. Rupesh Kumar Dewang and A. K. Singh. 2015. Identification of Fake Reviews Using New Set of Lexical and Syntactic Features. In Proc. of the Sixth International Conference on Computer and Communication Technology (ICCCT '15), 2015, pp. 115–119.

[19]. C. Zhao, W. Song, L. Liu, C. Du and X. Zhao. Research on Author Identification Based on Deep Syntactic Features. In Proc. of the 2017 10th International Symposium on Computational Intelligence and Design (ISCID), 2017, pp. 276-279.

[20]. K. Avetisyan and T. Ghukasyan. Word embeddings for the armenian language: intrinsic and extrinsic evaluation. Bulletin of the Russian-Armenian University: Physico-Mathematical and Natural Sciences, no. 1, 2019, pp. 59-72.

[21]. Gishamer Flurin. Using Hashtags and POS-Tags for Author Profiling. Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR Workshop Proceedings, vol. 2125, 2019.

Информация об авторах / Information about authors

Ева Максимовна ЕШИЛБАШЯН – студентка магистратуры по направлению машинного обучения факультета прикладной математики и информатики. Занимается обработкой естественного языка.

Yeva Maksimovna YESHILBASHIAN is a student of Machine Learning master's degree programme. Deals with natural language processing.

Ариана Арменовна АСАТРЯН – магистрант кафедры математической кибернетики. Научные интересы – интеллектуальный анализ текстов и машинное обучение.

Ariana Armenovna ASATRYAN is a master student at the Department of Mathematical Cybernetics. Her research interests include intelligent text analysis and machine learning.

Цолак Гукасович ГУКАСЯН является аспирантом кафедры системного программирования. Его научные интересы включают обработку естественного языка, машинное обучение.

Tsolak Gukasovitch GHUKASYAN is a postgraduate student of the Department of System Programming. His research interests include natural language processing, machine learning.

Приложение А. Список использованных признаков.

Уровень	Группа	Признаки
символы	пунктуация	• знаки пунктуации, их частота, например: ○ наличие знака пунктуации ○ #[знаки пунктуации] / #[слова] для всех знаков вместе, а также для каждого отдельно • комбинации пунктуации и пробельных символов, например: ○ наличие пробела перед знаком пунктуации ○ наличие пробела после знака пунктуации ○ #[наличие пробела перед знаком пунктуации] / #[знаки пунктуации] ○ #[наличие пробела после знака пунктуации] / #[знаки пунктуации] ○ вышеперечисленное для каждого знака пунктуации отдельно
	общие	• суффиксы: ○ наличие конкретного суффикса ○ #[слова с конкретным суффиксом] / #[слова] • префиксы: ○ наличие конкретного префикса ○ #[слова с конкретным префиксом] / #[слова]
слова	общие	○ частота слов: ○ наличие длинных слов ○ #[длинные слова] / #[слова] ○ наличие редких слов ○ #[редкие слова] / #[слова] ○ средняя частота используемых слов ○ наличие терминов на латинском/кириллице ○ наличие жаргона/неформальных выражений ○ написание именованных сущностей [1]
	сокращения	• использование сокращения вместо полной формы (например, թ. вместо թվական, թ-ի вместо թվականի) • стиль склонения сокращений (например, թ. vs թ.-ի vs թ-ի) • сокращения пишутся с маленькой буквы (например, բուհ, բnh) • сокращения пишутся с большой буквы (например, ԲՈՒՀ, ԲՈՀ) • наличие разделителя в сокращениях (դդ. vs դ.դ. или տտ. vs տ.տ.)
	числительные	• написание количественных (например, տասը հազար vs 10000) • написание порядковых (например, երկրորդ vs ii vs II vs 2-րդ vs 2րդ) • формат написания дат (например, 12\18\10 vs 12/18/10 vs 12-18-10 vs 12.18.10 vs 12\18\2010 vs 12/18/2010 vs 12-18-2010 vs 12.18.2010 vs 18 դեկտեմբերի, 2010)

	н-граммы	• биграммы: ○ наличие биграммы с низкой IDF ○ #[биграммы с низкой IDF] / (#[слов] - 1) • триграммы: ○ наличие траграммы с низкой IDF ○ #[триграммы с низкой IDF] / (#[слов] - 2)
Предло-жения	общие	• количество предложений • средняя длина предложений: ○ среднее число слов в предложении ○ среднее число символов в предложении • максимальная длина предложений: ○ максимальное число слов в предложении ○ максимальное число символов в предложении • минимальная длина предложений: ○ минимальное число слов в предложении ○ минимальное число символов в предложении • длинные предложения: ○ наличие длинных предложений ○ #[длинные предложения (с точки зрения числа слов)] / #[предложения] ○ #[длинные предложения (с точки зрения числа символов)] / #[предложения] • короткие предложения: ○ наличие коротких предложений ○ #[короткие предложения (с точки зрения числа слов)] / #[предложения] ○ #[короткие предложения (с точки зрения числа символов)] / #[предложения]
	морфологические	• части речи: ○ #[слова с этой частью речи] / #[слова] ○ наличие последовательности 2-х конкретных частей речи ○ #[последовательность 2-х конкретных частей речи] / (#[слова] - 1) ○ наличие последовательности 3-х конкретных частей речи ○ #[последовательность 3-х конкретных частей речи] / (#[слова] - 2)
	грамматические	• синтаксические связи, их частота ○ наличие синтаксической связи в параграфе (за исключением частых связей, которые повсеместны) ○ #[предложения, где присутствует синтаксическая связь] / #[предложения] • частота простых предложений: ○ наличие простых предложений ○ #[простые предложения] / #[предложения] • частота сложных предложений : ○ наличие сложных предложений ○ #[сложные предложения] / #[предложения]

Параграфы	читабельность	• Flesch reading ease • SMOG grade • Flesch-Kincaid grade • Coleman-Liau index • automated readability index • Dale-Chall readability score • difficult words • Linsear write formula • Gunning fog