О применимости γ-классификатора к распознаванию однородности текстов на славянских языках

А. А. Косимов, email: abdunabi_kbtut@mail.ru [1]

[1] Таджикский технический университет имени академика М.С.Осими

Аннотация. *В исследованиях Р.Грея и К.Аткинсона [1] посредством статистического анализа родственных слов, У.Чанга, Ч.Кэткарта, Д.Холла и А.Гарретта [2] с помощью статистического моделирования и А.С.Касьяна и А.В.Дыбо [3] на основе лексикостатистической классификации помимо обсуждения исторических вопросов представлены генеологические деревья, отражающие как родство, так и дивергенцию современных славянских языков. Таких деревьев достаточно много, они сходны в общих чертах и различны в небольших деталях, см. например, [3, 4]. Ареал прежде единого языка ныне разделился на три группы – восточную в составе белорусского, русского и украинского языков, западную - из чешского, словацкого, польского, кашубского и лужицких языков и южную, состоящую из болгарского, македонского, сербо-хорватского и словенского языков. В статье на примере случайно сформированной модельной коллекции из 26 текстов на 13 языках (по 2 произведения от каждого языка) устанавливается применимость γ-классификатора для автоматического распознавания принадлежности текстов той или иной группе славянских языков на основе частотности универсального для все языков набора латинских символов. Математическая модель γ-классификатора представляется в виде триады, составленной из цифрового портрета (ЦП) текста - распределения в тексте частотности латинских символьных униграмм; формулы для вычисления расстояний между ЦП текстами и алгоритма машинного обучения, реализующего гипотезу "однородности" произведений из одной группы языков и "неоднородности" произведений, принадлежащих разным группам языков. Настройка алгоритма, использующего таблицу парных расстояний между всеми произведениями модельной коллекции, осуществлялась путем подбора оптимального значения вещественного параметра γ, минимизирующего число ошибок нарушения гипотезы "однородности". Обученный на текстах модельной коллекции γ-классификатор показал 86%-ю точность в распознавании языков произведений. Для тестирования классификатора были выбраны 3 дополнительных случайных текста,*

по одному тексту для трёх разных групп славянских языков. Методом ближайшего (по расстоянию) соседа все новые тексты подтвердили свою однородность с соответствующими парами одноязычных произведений, тем самым и однородность с соответствующей группой славянских языков.

Ключевые слова: *текст, язык, славяне, алфавит, универсальный набор латинских символов, частотность, униграмм, цифровой портрет текста, классификатор, обучение, распознавания, группы языков, оценка эффективности, тестирование классификатора.*

Введение

Состояние работ по применение различных классификаторов, прежде всего методов нейронных сетей и машины опорных векторов, подробно описано в монографии [5]. В настоящей работе на примере модельной случайно сформированной коллекции из 26 произведений на 13 славянских языках (по 2 произведения от каждого языка) решаются две задачи:

- путем подбора вещественного параметра γ настроить так называемый γ-классификатор, по возможности, для безошибочного распознавания принадлежности текстов соответствующей одной из трёх групп языков;
- для трёх дополнительных случайно выбранных произведений, принадлежащих различным группам, проверить правильность работы настроенного классификатора.

Решение задач основано на применении γ-классификатора – математической триады, первым компонентом которой является цифровой портрет (ЦП) текста – распределение в тексте частотности буквенных униграмм; вторым компонентом служит формула для вычисления расстояний между ЦП текстов и третьим – алгоритм машинного обучения, реализующий гипотезу “однородности” произведений, принадлежащих одной группе языков, и “неоднородности” произведений, принадлежащих разным группам языков. Настройка алгоритма, использующего таблицу парных расстояний между всеми произведениями модельной коллекции, заключалась в определении полуинтервала значений вещественного параметра γ, на которых минимизируется ошибка нарушения гипотезы “однородности”. Обученный на текстах модельной коллекции γ-классификатор тестируется на предмет правильного отнесения случайного текста группе “однородных” с ним произведений.

Прежде чем переходить к изучению задач, напомним основные понятия, связанные с компонентами триады.

1. Модельная коллекция текстов С,

собранная случайным образом, представляет три группы славянских языков, причём от каждого языка по два произведения. В приводимом далее списке элементов коллекции С указываются имя автора, название его сочинения на родном языке и в скобках – используемый алфавит, аббревиатура сочинения и его размеры в количестве слов:

а) в восточнославянской группе

на белорусском языке:

Л.Станислав “Салярыс, часть 1” (кир., be_1, 8497 слов);

С.Давидович (Be): “Дзед-кіёк” (кир., be_2, 1935 слов);

на русском языке:

М.А.Шолохов (Ru): “Судьба человека” (кир., ru_1, 10891 слов);

Ф.А.Абрамов (Ru): “Алька” (кир., ru_2, 15668 слов);

на украинском языке:

В.Л.Кашин “Готується вбивство” (кир., uk1, 23771 слов);

М.Циба (Uk): “Акванавти, або Золота жила” (кир., uk_2, 20150 слов);

б) в западнославянской группе

на польском языке:

R.M.Wegner “Jeszcze może załopotać, часть 1” (лат., pl1, 10601 слов);

R.M.Wegner “Jeszcze może załopotać, часть 2” (лат., pl2, 9670 слов);

на чешском языке:

S.Lem “K Mrakům Magellanovým” (лат., cs1, 17552 слов);

B.S.R.Jordan “Bouře přichází” (лат., cs2, 17439 слов);

на словацком языке:

I.A.Jefremov “Na hranici Oekumeny” (лат., sv1, 13534 слов);

J.Jesenský “Demokrati” (лат., sv2, 17113 слов);

на кашубском языке:

D.Pioch “Biuletin Radzëznë Kaszëbsczégò Jãzëka” (лат., ks1, 12070 слов);

E.Breza “Prymas z Kaszub” (лат., ks2, 16871 слов);

в) в южнославянской группе:

на болгарском языке:

Н.Райнов “Неволя и богатство” (кир., bo1, 2565 слов);

Б.Джим (Bo): “Фурията на принцепса, глава 1” (кир., bo_2, 2491 слов);

на боснийском языке:

И.Асимов “Немезис” (кир., bs1, 20035 слов);

Д.Вейнс “Mjесечев мољац” (кир., bs2, 10443 слов);

на сербском языке:

А.Кларк “Напеви далеке Земље” (кир., se1, 11129 слов);

Р.Л.Стивенсон “Црна стрела” (кир., se2, 15028 слов);

на словенском языке:

M.Hudnik “Kakor Kartagina” (лат., sl1, 14626 слов);

I.Koprivec “Josip Vidmar v oieh svojih sodobnikov” (лат., sl2, 16985 слов);

на македонском языке:

В.Тоциновски “Кочо Рацин - наша творечка и етичка мерка” (кир., mk1, 9047 слов);

Г.Прличев “Сердарот” (кир., mk2, 9478 слов);

на хорватском языке:

I.M.Andrić “Pročitani Pisci (Eseji i prikazi)” (лат., xr1, 26221 слов);

M.Lovrak “Vlak U Snijegu” (лат., xr2, 10522 слов).

2. Цифровой портрет произведений

В качестве элементов количественного образа произведений нами используются буквенные униграммы. Поскольку для славянских языков нет единого буквенного алфавита (в указанном списке 14 произведений на основе кириллического алфавита и 12 – на основе латинского), мы осуществляем предобработку алфавитов таким образом, чтобы выделить в них унифицированный набор символов. Среди 14 аналогов кириллических алфавитов общими оказались 26 букв: – “а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ю, я”; между тем для 12 аналогов латинского алфавита – тоже 26 букв, но уже следующие “a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z”. Из этих двух алфавитов был сформирован искусственный общий для всех текстов алфавит из 22 символов “a, b, c, d, e, f, g, i, j, k, l, m, n, o, p, r, t, u, v, x, y, z“, учитывающих сходных по написанию и по звучанию символы.

Теперь, когда хотя бы формально, все тексты описываются одним и тем же набором из 22 латинских символов, введем следующее

Определение 1. Цифровым портретом (ЦП) какого-либо текста T на славянском языке будем называть распределение в нём частотности упомянутых 22 латинских символов.

ЦП текста T записывается в табличном виде:

N : 1 2 … 22

P : p_1 p_2 … p_{22} ,

в котором первая строка – номера символов, расположенных в алфавитном порядке, а вторая □ относительные частоты встречаемости символов в тексте T, причём $\sum_{k=1}^{22} p_k = 1$.

Цифровой портрет представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{s} p_k \quad (s = 1,..., 22). \tag{1}$$

3. Расстояния между цифровыми портретами текстов

Пусть T_1 , T_2 – произвольная пара текстов из С, характеризуемых на основе единого символьного алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \tag{2}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1,2,$ и $(s = 1,..., 22)$.

Определение 2. Расстоянием между текстами T_1 и T_2 называется положительное число $\rho(T_1, T_2)$, определяемое формулой

$$\rho(T_1, T_2) = \sqrt{\frac{22}{2}} \max_s \left| F^{(1)}(s) - F^{(2)}(s) \right| \tag{3}$$

4. Гипотеза Н “однородности” произведений

Она привлекается для того чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания однородных групп произведений. Её мы формулируем в следующем виде.

ГИПОТЕЗА Н. Любая пара произведений из одной и той же группы славянских языков “однородна”, а из разных групп “не однородна”.

Говоря об “однородности” произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

5. Математическая модель Н-гипотезы

Пусть γ - некоторое положительное число.

Определение 3. Тексты T_1, T_2 называются γ-однородными (принадлежащими одной и той же группе славянских языков), если

$$\rho(T_1, T_2) \le \gamma, \tag{4}$$

и γ-неоднородными (принадлежащими разным группам славянских языков), если

$$\rho(T_1, T_2) > \gamma. \tag{5}$$

Неравенства (4) и (5) являются математической интерпретацией (моделью) гипотезы Н.

Определение 4. γ - классификатор – это зависящий от одного вещественного параметра γ алгоритм принятия решения об отнесении

пары текстов T_1 и T_2 к одной или двум разным группам славянским языкам.

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов одной группе языков в рамках математической модели означает справедливость неравенства (4), а двум разным группам – справедливость неравенства (5). Гипотеза H может нарушаться для каких-то пар текстов одной и той же группе языков в случае, когда вместо неравенства (4) имеет место неравенство (5), а также в случае, когда какие-то два текста из разных групп удовлетворяют неравенству (4) вместо того, чтобы выполнялось неравенство (5).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы H одновременно в двух случаях: невыполнения неравенства "однородности" в случае двух текстов, принадлежащих одной группе, и невыполнения неравенства "неоднородности" в случае двух текстов, принадлежащих разным группам. Тогда для фиксированного γ показатель выполнения гипотезы будем определять величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma) / L, \qquad (6)$$

где L - число взаимных расстояний между всеми парами текстов из коллекции C (в нашем случае $L = C_{26}^2 = 325$). Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём π = 0, если $\tau = L(= 325)$, и π = 1, если τ = 0. В первом случае гипотезу H следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию принадлежности пары произведений одной или же различным группам. Алгоритм настройки классификатора приведен в [6].

6. Предварительные результаты на примере модельной коллекции C

приведены далее путём последовательного выполнения следующих операций:

- вычисления цифровых портретов (частотности букв 22 общих латинских символов) для всех 26 произведений модельной коллекции С;

- вычисления по формулам (1), (2) и (3) 325 парных расстояний $\rho(T_1, T_2)$ между произведениями коллекции С (результаты расчетов приведены в последующей таблице):

Таблица 1

Расстояния между текстами коллекции С

Тексты		Восточнославянская подгруппа						Западнославянская подгруппа								Южнославянская подгруппа											
		be1	be2	ru1	ru2	uk1	uk2	pl1	pl2	cs1	cs2	sv1	sv2	ks1	ks2	bo1	bo2	bs1	bs2	se1	se2	sl1	sl2	mk1	mk2	xr1	xr2
Вост.	be1																										
	be2	0.13																									
	ru1	0.36	0.45																								
	ru2	0.27	0.35	0.09																							
	uk1	0.39	0.51	0.17	0.25																						
	uk2	0.36	0.47	0.13	0.21	0.04																					
Запад.	pl1	0.36	0.39	0.29	0.27	0.26	0.24																				
	pl2	0.33	0.36	0.28	0.26	0.28	0.25	0.03																			
	cs1	0.40	0.43	0.15	0.14	0.24	0.21	0.25	0.27																		
	cs2	0.34	0.37	0.13	0.12	0.27	0.24	0.21	0.23	0.06																	
	sv1	0.30	0.33	0.15	0.12	0.28	0.25	0.22	0.22	0.11	0.07																
	sv2	0.29	0.32	0.14	0.07	0.30	0.26	0.24	0.23	0.13	0.09	0.05															
	ks1	0.37	0.40	0.31	0.29	0.30	0.26	0.11	0.09	0.25	0.23	0.20	0.22														
	ks2	0.37	0.40	0.25	0.23	0.28	0.24	0.04	0.04	0.25	0.22	0.18	0.20	0.09													
Южн.	bo1	0.20	0.27	0.28	0.22	0.36	0.33	0.35	0.34	0.35	0.31	0.24	0.23	0.34	0.31												
	bo2	0.20	0.29	0.23	0.17	0.32	0.28	0.31	0.30	0.30	0.26	0.18	0.18	0.30	0.26	0.13											
	bs1	0.22	0.28	0.30	0.24	0.37	0.34	0.37	0.36	0.37	0.33	0.26	0.25	0.39	0.33	0.09	0.11										
	bs2	0.27	0.34	0.25	0.19	0.32	0.28	0.36	0.35	0.32	0.28	0.21	0.20	0.37	0.32	0.11	0.09	0.10									
	se1	0.23	0.30	0.27	0.21	0.35	0.31	0.36	0.35	0.34	0.30	0.22	0.22	0.38	0.32	0.09	0.09	0.05	0.08								
	se2	0.27	0.32	0.30	0.24	0.37	0.34	0.35	0.34	0.37	0.33	0.26	0.25	0.37	0.31	0.11	0.09	0.06	0.06	0.05							
	sl1	0.30	0.38	0.31	0.25	0.27	0.26	0.36	0.35	0.33	0.30	0.25	0.23	0.36	0.32	0.14	0.14	0.10	0.11	0.09	0.11						
	sl2	0.35	0.43	0.29	0.23	0.27	0.26	0.31	0.30	0.31	0.28	0.23	0.21	0.33	0.27	0.18	0.17	0.15	0.10	0.14	0.16	0.05					
	mk1	0.22	0.31	0.23	0.17	0.30	0.27	0.35	0.34	0.30	0.26	0.20	0.18	0.35	0.31	0.21	0.08	0.17	0.13	0.14	0.17	0.20	0.19				
	mk2	0.16	0.23	0.29	0.23	0.40	0.36	0.37	0.36	0.36	0.32	0.25	0.24	0.39	0.33	0.09	0.09	0.06	0.11	0.08	0.12	0.15	0.20	0.13			
	xr1	0.31	0.39	0.35	0.29	0.30	0.27	0.36	0.35	0.37	0.33	0.28	0.27	0.38	0.32	0.14	0.13	0.15	0.12	0.12	0.15	0.07	0.05	0.20	0.17		
	xr2	0.24	0.29	0.40	0.33	0.35	0.32	0.38	0.37	0.41	0.38	0.33	0.31	0.40	0.34	0.12	0.18	0.14	0.16	0.15	0.12	0.09	0.14	0.26	0.13	0.11	

- вычисление с помощью алгоритма настройки γ-классификатора [6] оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы H достигает минимального значения и, следовательно, величина π показателя выполнения гипотезы H принимает максимальное значения.

На данных таблицы 1 вычислен оптимальный полуинтервал значений γ

$\gamma^{опт} \in [0.2142; 0.2160)$

В соответствии с определением 3 это значит, что если расстояние $\rho(T_1, T_2)$ между двумя текстами не превосходит значения $\gamma^{опт} < 0.2160$, то пара текстов принадлежат одной и той же группе языков; если же $\rho(T_1, T_2)$ превосходит 0.2160, то принадлежат разным языкам.

Минимальное число нарушений оказалось равным τ = 45. В таблице 1 ячейки нарушения гипотезы (4) “однородности” отмечены слабо серым цветом, а гипотезы (5) “неоднородности” серым цветом.

Теперь остается вычислить эффективность π классификатора по формуле (6):

$\pi = 1 - \tau(\gamma^{опт}) / L = 0.86$

7. Тестирование классификатора

После того как за счёт выбора оптимального значения γ произошла настройка классификатора и был отработан алгоритм, который в 86 случаях из 100 правильно соотносил элементы модельной коллекции к соответственной группе славянских языков, возникает естественный вопрос, а каковы будут результаты раскладки уже других славянских текстов, не входящих в коллекцию, по тем же самым трем языковым группам.

Для тестирования классификатора выбраны случайным образом 3 текста:

на украинском языке (Uk) - В.П.Бережной “Homo Novus” (кир., Text_Uk, 5768 слов);

на польском языке (Pl) - A.Szklarski “Tomek wśród łowców głów” (лат.,Text_Pl, 13635 слов);

на больгарском языке (Bo) - А.Каралийчев “Гулчечек” (кир., Text_Bo, 2436 слов).

Для каждого произведения так же, как это было сделано для всех текстов модельной коллекции, построены ЦП на основе единого набора из 22 латинских символах. После чего по формуле (3) вычислены расстояния до всех 26 элементов модельной коллекции. Результаты показаны в таблице.

Таблица 2

Расстояния между текстами коллекции С и тремя случайно выбранными произведениями

Тексты		**Text_Uk**	**Text_Pl**	**Text_Bo**
Восточная группа	**be1**	0.3421	0.3432	0.2031
	be2	0.4490	0.3742	0.2926
	ru1	0.1034	0.2699	0.2131
	ru2	0.1896	0.2517	0.1515
	uk1	0.0714	0.2398	0.3297
	uk2	**0.0511**	0.2190	0.2912
Западная группа	**pl1**	0.1612	0.1916	0.2013
	pl2	0.1791	0.2030	0.1836
	cs1	0.1856	0.2070	0.2844
	cs2	0.2125	0.1745	0.2445
	sv1	0.2238	0.2010	0.2090
	sv2	0.2391	0.2162	0.1619
	ks1	0.2347	0.1271	0.3233
	ks2	0.2158	**0.0862**	0.2946
Южная группа	**bo1**	0.3014	0.3389	0.1064

Тексты		Text_Uk	Text_Pl	Text_Bo
	bo2	0.2578	0.2901	**0.0510**
	bs1	0.3165	0.3521	0.1331
	bs2	0.2560	0.3386	0.1035
	se1	0.2918	0.3407	0.1115
	se2	0.3129	0.3303	0.1086
	sl1	0.2192	0.3418	0.1403
	sl2	0.2049	0.2971	0.1822
	mk1	0.2458	0.3232	0.1206
	mk2	0.3350	0.3500	0.0936
	xr1	0.2441	0.3419	0.1533
	xr2	0.2921	0.3661	0.2013

В ячейках таблицы, на пересечении столбцов и строк, приводятся значения расстояний между текстами. В первых трех столбцах ближайшими соседями текстов Text_Uk, Text_Pl и Text_Bo являются соответственно uk2, ks2 и bo2 на расстояниях соответственно 0.0511, 0.0862 и 0.0510 (в таблице отмечены серым цветом). Полученный результат показывает, что по методу ближайшего соседа три случайно выбранных произведения распределяются как раз по тем группам языков, которым они сами принадлежат.

Заключение

Итак, γ-классификатор с фиксированным значением $\gamma = \gamma^{опт}$ на случайных выборках текстов с цифровыми портретами на основе частотности 22 латинских символов подтвердил 86%-ную статистическую способность к распознаванию групп произведений на славянских языках. В свою очередь, метод ближайшего соседа показал возможность безошибочного распределения дополнительных славянских произведений по восточной, западной и южной группам славянских языков.

Список литературы

1. Russell D., Gray and Quentin D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin // Nature: журнал. – Великобритания: Nature Publishing Group, 2003. – Т.426. –№6965. – С. 435-439.

2. Chang, W. «Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis (= Филогенетический анализ, связанный с предками, подтверждает гипотезу индоевропейских степей)» Language / W. Chang, Ch. Cathcart, D. Hall, A. Garrett // Volume

91. – Number 1. –March 2015. – PP. 194-244 (Article). Published by Linguistic Society of America.

3. Kassian A., Dybo A. Supplementary Information 2: Linguistics: Datasets; Methods; Results (в статье Kushniarevich A., Utevska O., Chuhryaeva M., Agdzhoyan A., Dibirova K., Uktveryte I. et al. (2015) Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. PLoS ONE 10(9): e0135820. https://doi.org/10.1371/journal.pone.0135820).

4. Генеалогическое дерево славянских языков – Википедия [Электронный ресурс]. – Режим доступа: http://900igr.net/kartinka/biologija/tema-proiskhozhdenie-jazykov-127785/5.-genealogicheskoe-drevo-slavjanskikh-jazykov-2.html

5. Романов, А.С. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста / А.С. Романов, А.А. Шелупанов, Р.В. Мещеряков // – В-Спектр: Томск, – 2011. – 188 с.

6. Усманов, З.Д. Алгоритм настройки кластеризатора дискретных случайных величин / З.Д. Усманов // ДАН РТ. – 2017. – Т.60. – № 9. – С. 392-397.