Ф.Н. Соловьев, А. М. Чеповский

Расширение модели идентификации языка коротких текстов[1]

Аннотация. В статье рассмотрена проблема автоматической идентификации естественного языка текста и наиболее полная известная нам ее модель. Предлагается расширение модели на новые кириллические языки малых народов России.

Ключевые слова: статистическая модель языка, идентификация естественного языка текста, языки малых народов России.

Введение

В настоящее время все более востребованы средства автоматического анализа текстов, призванные упростить и автоматизировать работу с все возрастающими объемами информации. Такие средства находят свое применение при работе с информационными и коммуникационными системами, системами электронного документооборота, анализа потоков данных и пр.

Важной задачей автоматического анализа текстов является автоматическое определение языка текста. Успешное ее решение позволяет при дальнейшем анализе текста учитывать структуру и особенности языка, на котором он написан.

В современном мире формат коротких текстовых сообщений при коммуникации становится все более популярным. Помимо электронной почты и службы коротких текстовых сообщений (SMS), широкую популярность приобрели различные социальные сети (такие как Твиттер, Инстаграм, Фейсбук, Вконтакте, Одноклассники). Многие Интернет-сервисы, размещающие информационные статьи, теперь включают возможность комментирования публикуемых материалов.

Известно множество различных методов идентификации языка длинных (сотни символов и более) образцов текстов, однако распознавание очень коротких строк (единицы и десятки символов) по-прежнему представляет сложность для систем автоматической обработки информации.

Характерно, что данная задача рассматривается как крайне актуальная для западных исследователей. В частности, решению данной проблемы посвящены работы [3, 4], в которых, однако, набор определяемых языков невелик (порядка 10 языков). В отечественной работе [1] анализируется достаточно широкий набор языков (59 языков и 62 пары язык-система письма).

В настоящей работе мы используем подход [1], расширяя множество распознаваемых языков новыми кириллическими языками малых народов России, полученными в работе [2]. Мы выбираем такое подмножество добавляемых языков, которое позволяет, не сильно уменьшая качество определения исходных языков, получить более полную модель.

1. Вероятностная языковая модель

Следуя работе [1] мы используем вероятностный подход для решения задачи определе-

[1]Работа выполнена при поддержке РФФИ, гранты № 16-29-09546 офи_м и № 16-07-00641.

ния языка текста, опирающийся на модель Байесовского классификатора. Опишем основные моменты применяемого подхода.

Пусть имеется бесконечное счетное множество объектов $X = \{x_1, \dots, x_i, \dots\}$ и конечное множество классов $Y = \{y_1, \dots, y_K\}$. Рассмотрим условное распределение:

$$P(y_i|x_j) \quad (1)$$

Подход Байесовского классификатора состоит в том, что объекту x_j ставится в соответствие класс y_i максимизирующий вероятность (1):

$$y(x_j) = argmax_{y_i} P(y_i|x_j) = argmax_{y_i} \frac{P(y_i)P(x_j|y_i)}{P(x_j)} = \\ = argmax_{y_i} P(y_i)P(x_j|y_i) = argmax_{y_i} P(y_i)P_{y_i}(x_j). \quad (2)$$

В нашем случае Y – множество языков, а X – множество текстов. Мы считаем, что все языки равновероятны, а значит соотношение (2) можно в нашем случае переписать как

$$y(x_j) = argmax_{y_i} P_{y_i}(x_j). \quad (3)$$

Пусть текст x есть последовательность символов $s = c_1 \dots c_m$, принадлежащих некоторому конечному алфавиту Σ. Тогда вероятностное распределение $P_y(s)$ для языка y можно переписать следующим образом:

$$P_y(s) = P_y(c_1 \dots c_m) = P_y(c_1) \prod_{k=2}^{m} P_y(c_k|c_{k-1} \dots c_1). \quad (4)$$

Полагая, что в соотношении (4) условная вероятность символа $P_{y_i}(c_k|c_{k-1} \dots c_1)$ зависит не более чем от $n-1$ предыдущих символов, получим приближение вероятности (4):

$$\tilde{P}_y(s) = \\ = P_y(c_{n-1}|c_{n-2} \dots c_1) \dots P_y(c_1) \prod_{k=n}^{m} P_y(c_k|c_{k-1} \dots c_{k-(n-1)}). \quad (5)$$

Вероятности $P_y(c_k|c_{k-1} \dots c_{k-(n-1)})$ аппроксимируются с помощью частот $f_y(c_{k-(n-1)} \dots c_k)$ встречаемости в текстах языка и сглаживания [5]:

$$P_y(c_n|c_{n-1} \dots c_1) = \begin{cases} \frac{f_y(c_1 \dots c_n)}{f_y(c_1 \dots c_{n-1})} * (1 - p_0), & f_y(c_1 \dots c_n) \neq 0, \quad f_y(c_1 \dots c_{n-1}) \geq \theta \\ p_0, & f_y(c_1 \dots c_n) = 0, \quad f_y(c_1 \dots c_{n-1}) \geq \theta \\ P_y(c_n|c_2 \dots c_{n-1}), & f_y(c_1 \dots c_{n-1}) < \theta \end{cases}, \quad (6)$$

где $P_y(c_n) = \begin{cases} f_y(c_n) * (1 - p_0), & f_y(c_n) \neq 0 \\ p_0, & f_y(c_n) = 0 \end{cases}$ (7)

p_o – константа, задающая пороговый уровень вероятности для подстрок с малой частотой появления;

θ – константа, определяющая пороговое значение глубины учета длин подстрок при вычислении вероятностей для подстроки.

Значения параметров p_o и θ определяются экспериментально.

Для фрагмента текста s может быть вычислена числовая оценка соответствия фрагмента языковой модели языка y. Пусть фрагмент текста s содержит n символов и может быть подсчитана в соответствии с формулами (6)-(7) вероятность $P_y(s)$ для данной строки. Тогда оценка соответствия этого фрагмента языку y определяется как:

$$E_y(s) = \frac{ln(P_y(s))}{n}. \quad (8)$$

Такая оценка согласно [6] представляет собой величину, стремящуюся к значению количества информации (по Шеннону) на символ текста при стремлении n к бесконечности и не зависит от длины текста. Дисперсия уменьшается с ростом длины фрагмента.

Оценка (8) соответствия языку y фрагментов текста сильно зависит от используемого другими языками алфавита и степени похожести языков на язык y.

В работе [1] утверждается, что если алфавиты языков не пересекаются с русским алфавитом (английский), то тексты на этих языках будут отсекаться с высокой точностью. Если же алфавиты языков пересекаются с русским алфавитом, то невозможно точно отделить тексты на другом языке при тривиальном пороговом отсечении согласно максимуму значения вероятности.

При увеличении длины фрагмента шансы отделения неизвестного языка с помощью порогового значения увеличиваются. Для отсечения неизвестных языков используется пороговое значение оценки (8), которое вычисляется отдельно для каждого языка и каждой длины фрагмента:

$$T_y = M[E_y] - k * \sigma[E_y], \quad (9)$$

где $M[E_y]$ – математическое ожидание оценки (8) соответствия для языка y;

$\sigma[E_y]$ – среднеквадратичное отклонение оценки (8) соответствия для языка y;

k– настраиваемый коэффициент порога отсечения; определяется экспериментально.

Текст, имеющий оценку (8) соответствия языку y ниже порога (9), не может быть отнесен к языку y и считается написанным на неизвестном языке.

Рассматриваемый набор текстов содержит тексты на языках малых народов России, не анализируемых в работе [1]. Будем называть множество языков, исследуемых в работе [2] *новыми*. В настоящей работе указанные тексты используются для расширения модели из [1]. Будем называть языки, рассматриваемые в [1], *исходными*, а модель классификации языка текста, описанную в [1], – просто *моделью*.

2. Расширение исходной модели языками малых народов России

В настоящей работе мы, опираясь на результаты работы [1], расширяем множество определяемых языков. В [2] была проведена значительная работа по составлению набора текстов на языках малых народов России, на основе текстов сайтов из сети Интернет, а также текстов социальной сети Вконтакте. Полученные тексты использовались нами для расширения модели [1]. В работе [2] использовалось API поисковой системы Яндекс для определения источников текстов, написанных на языках малых народов России. Ввиду зашумленности данных, в работе [2] использовался следующий подход. Для каждого источника анализировался отрывок текста на предмет соответствия русскому языку, поскольку для него имеются достаточно точные методики определения языка. Тексты, отнесенные на основе анализируемого отрывка текста к русскому языку, исключались из рассмотрения. Полученная выборка использовалась нами для расширения имеющегося механизма определения языка.

Тексты на новых языках были дополнительно отфильтрованы с помощью модели [1], на каждом из новых языков – в один файл, фильтруемый построчно. Отфильтровывались все строки короче 10 символов. Как правило, такие строки содержали сопроводительную к основному тексту информацию, зачастую техническую (даты, номера документов, комментариев и т.п.). Для каждой строки анализировалось 200 первых символов (в случае коротких строк – строка целиком) на предмет отнесения к одному из языков модели [1]. Отфильтровывались строки, отнесенные к русскому языку. Также были удалены все дублирующие строки.

Следуя методике тестирования, описанной в [1], в настоящей работе измеряются такие характеристики как точность, полнота и F-мера. Пусть n_0^y – тексты, написанные на языке y, а n_0^{-y} – тексты, написанные на остальных языках. Пусть n^y – тексты, в которых язык был определен как y, а n^{-y} – тексты, в которых был определен другой язык. Пусть n_+^y – тексты, в которых язык был *верно* определен как y, а n_-^y – тексты, в которых был *неверно* определен язык y. Тогда точность $\Pr(y)$, полнота $R(y)$ и F-мера $F(y)$ определения языка y определяются как

$$\Pr(y) = \frac{n_+^y}{n_+^y + n_-^y},$$
$$P(y) = \frac{n_+^y}{n_0^y}, \quad (10)$$
$$F(y) = \frac{2}{\Pr(y) + R(y)}.$$

Так же как и в [1] для тестирования тексты на каждом из языков объединялись в один текстовый файл для каждого языка. Полученный текстовый файл разбивался на тренировочную и тестовую части для обучения и тестирования модели. Однако, в отличие от работы [1], размер тестовой части фиксировался равным 20000 символов, вместо выделения 20% исходного текста. Такой подход призван снизить влияние размера исходного текстового файла на точность тестирования, одновременно позволяя использовать как можно больший объем текста для обучения модели.

Для сравнения модели, обученной на исходных текстах и модели, обученной на исходных текстах, дополненных новыми текстами, мы измеряем точность, полноту и F-меру на множестве текстов, включающем только исходные языки.

Отдельно была протестирована модель, обученная на наборе текстов на исходных и новых языках, используя тестовый набор текстов, написанных исключительно на новых языках, с тем, чтобы получить оценку качества опреде-

ления новых языков моделью, обученной на расширенном множестве языков.

При тестировании модели, обученной на расширенном на все 30 новых языков множестве, было установлено, что многие тексты на новых языках достаточно зашумлены, о чем можно судить по снижающейся точности определения русского языка.

Для того чтобы повысить точность модели, множество новых языков было сокращено до 10, оставляя те языки, для которых при тестировании на текстах на этих языках количество раз ложного определения русского языка минимально. Все 30 новых языков включают:

Адыгейский, Алтайский, Калмыцкий, Кашмири, Коми-пермяцкий, Корякский, Кумыкский, Мансийский, Мокшанский, Нанайский, Ненецкий, Нивхский, Ногайский, Тувинский, Удмуртский, Хакасский, Цахурский, Чувашский, Чукотский, Шорский, Эвенкийский, Эрзянский, Якутский, Удинский, Рутульский, Коми-зырянский, Бурятский, Марийский, Татский, Карагасский.

Из них было отобрано 10 следующих:

Адыгейский, Кашмири, Корякский, Кумыкский, Ненецкий, Ногайский, Шорский, Удинский, Татский, Карагасский.

В Табл. 1 и Табл. 2 представлены зависимости F-меры от длины текста, для которого определяется язык, а также от величины коэффициента отсечения k. Значение параметра выбирается из множества {2, 3}. В таблицах приводятся значения F-меры, вычисленные для тестовых фрагментов текста размером в 30 и 60 символов.

Табл. 1. F-мера, вычисленная на тестовых текстах на некоторых языках из исходного набора языков при обучении модели со значениями параметров k=2,3, длине текста 30 и 60 символов на исходном наборе, наборе, расширенном на 10 языков и наборе, расширенном на 30 языков

	Набор	**Исх.**	**Нов.(30)**	**Нов.(10)**	**Исх.**	**Нов.(30)**	**Нов.(10)**
	k	**2**	**2**	**2**	**3**	**3**	**3**
Язык - Письменность	**Длина текста**						
Белорусский - cyr	**30**	83,76	82,41	82,57	**91,84**	87,36	88,8
Белорусский - cyr	**60**	75,31	76,48	78,2	88,33	88,01	**88,64**
Итальянский - lat	**30**	**94,08**	93,57	93,35	93,97	93,63	93,59
Итальянский - lat	**60**	96,48	96,69	96,71	**97,85**	97,55	97,75
Кабардино-черкесский - cyr	**30**	97,55	90,67	91,08	**98,89**	90,13	90,63
Кабардино-черкесский - cyr	**60**	97,44	94,13	94,97	**99,09**	94,84	95,92
Карачаево-балкарский - cyr	**30**	96,95	87,55	86,47	**97,08**	87,74	86,49
Карачаево-балкарский - cyr	**60**	98,84	90,72	90,83	**99,5**	91,02	91,08
Киргизский - cyr	**30**	98,85	94,44	95,05	**98,95**	94,21	95,13
Киргизский - cyr	**60**	**99,7**	98,86	98,86	**99,7**	98,86	98,96
Монгольский - cyr	**30**	98,84	96,18	98,73	**99,55**	96,43	**99,55**
Монгольский - cyr	**60**	99,09	98,34	98,73	99,65	98,81	**99,8**
Осетинский - cyr	**30**	66,84	66,89	67,72	85,5	84,58	**85,63**
Осетинский - cyr	**60**	50,63	52,4	52,62	74,69	75,7	**75,93**
Польский - lat	**30**	99,14	99,35	98,94	99,9	**99,95**	**99,95**
Польский - lat	**60**	99,19	99,65	99,4	99,9	**99,95**	99,9
Русский - cyr	**30**	88,5	66,53	86,23	**89,08**	65,39	87,19
Русский - cyr	**60**	91,25	79,73	90,55	**95,7**	79,59	94,59
Словенский - lat	**30**	96,26	95,81	95,96	**96,42**	96,08	96,23
Словенский - lat	**60**	99,05	98,84	98,84	**99,45**	99,4	99,35
Татарский - cyr	**30**	95,12	95,01	95,59	**96,43**	95,09	95,89
Татарский - cyr	**60**	96,48	99,05	99,25	99,19	99,3	**99,5**
Турецкий - lat	**30**	99,45	99,45	99,7	99,75	99,6	**99,9**
Турецкий - lat	**60**	99,75	**100**	99,85	**100**	**100**	**100**
Украинский - cyr	**30**	**97,8**	94,24	93,91	97,52	93,39	93,79
Украинский - cyr	**60**	99,75	96,55	95,9	**99,8**	96,22	95,71

Табл. 2. F-мера, вычисленная на тестовых текстах на некоторых языках из 10 новых языков при обучении модели со значениями параметров k=2,3, длине текста 30 и 60 символов на исходном наборе, наборе, расширенном на 10 языков и наборе, расширенном на 30 языков

	Набор	**Нов.(30)**	**Нов.(10)**	**Нов.(30)**	**Нов.(10)**
	k	**2**	**2**	**3**	**3**
Язык - Письменность	**Длина текста**				
Адыгейский - cyr	**30**	76,54	77,1	80,25	**81,85**
Адыгейский - cyr	**60**	79,76	83,1	86,99	**88,9**
Корякский - cyr	**30**	53,61	61,31	58,04	**67,77**
Корякский - cyr	**60**	51,14	59,95	55,8	**67,2**
Кумыкский - cyr	**30**	32,33	33,4	32,62	**33,99**
Кумыкский - cyr	**60**	30,18	28,41	**30,22**	28,87
Марийский - cyr	**30**	93,22	**94,94**	93,08	94,61
Марийский - cyr	**60**	96,95	97,06	96,71	**97,54**
Ногайский - cyr	**30**	0,4556	**54,31**	0,1034	53,38
Ногайский - cyr	**60**	0,3849	55,53	nan	**60,17**

Исследовались модели, обученные на нескольких раличных множествах языков, а именно:

- множество исходных языков;
- множество исходных языков, расширенное текстами 30 новых языков (социальная сеть Вконтакте и страницы сети Интернет);
- множество исходных языков, расширенное текстами 10 новых языков (социальная сеть Вконтакте и страницы сети Интернет).

В Табл. 1 представлены значения F-меры при определении русского языка и некоторых других языков, взятых из набора исходных языков. Размер тестируемого участка текста варьируется от 10 до 60 символов с шагом в 10 символов. Положение участка текста выбирается случайным образом, но так, чтобы участок начинался с начала некоторого слова в тексте, как и в работе [1]. Мы не приводим значения F-меры для всех рассматриваемых языков, поскольку для большинства языков рассматриваемые значения мер практически не изменяются по сравнению с моделью, обученной на исходном множестве языков. В целом, значение F-меры остается примерно на прежних значениях и различия находятся в пределах 1%. Мы, однако, приводим результаты для некоторых из языков, для которых различие значений исследуемых мер достаточно заметно. Из Табл. 1 можно видеть, что при определении Украинского, Татарского, Кабардино-Черкесского, Киргизского, Карачаево-Балкарского, Белорусского языков F-мера падает при добавлении новых языков на 3-5%.

Также, можно видеть, что для Монгольского и Осетинского языков заметно ниже F-мера для модели, обученной на 30 новых языках, чем для модели, обученной на 10 новых языках.

В Табл. 2 приведены значения F-меры при определении некоторых языков, взятых из набора текстов, расширенных на 10 новых языков, полученные на моделях, обученных на расширенном на 10 новых языков множестве текстов. Для Кумыкского языка наблюдаются совсем невысокие значения F-меры. Для Корякского и Ногайского значение также невелико, но уже достаточно приемлемо для автоматического определения языка. Для остальных семи языков значения F-меры уже для 30 символов достаточно высоки.

Заключение

В настоящей работе была получена модель определения языка, расширенная на 10 новых языков. Несмотря на некоторое снижение точности определения русского языка, полученная модель позволяет, сохраняя в целом качество определения исходных языков, достаточно эффективно определять язык текстов для добавляемых в модель языков. Падение качества распознавания при расширении модели, связанного с добавлением 30 языков, может быть объяснено зашумленностью данных, полученных в ходе исследования [2].

Литература

1. Гусев С. В., Чеповский А. М. Модель для иденти- фикации естественного языка текста // Бизнес- информатика. 2011. №3 (17).
2. Зайдельман Л. Я., Крылова И. В., Орехов Б. В. Технология поиска и сбора в Интернете текстов на малых языках России // В кн.: Труды Международной научной конференции по физико-технической информатике (CPT2015). М.: Протвино. Институт физико-технической информатики, 2016. С. 179-181.
3. Vogel J., Tresner-Kirsch D. Robust language identification in short, noisy texts: Improvements to LIGA // Proceedings of the Third International Workshop on Mining Ubiquitous and Social Environments, 2012. - P. 43-50.
4. Carter S., Weerkamp W., Tsagkias M. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text // Language Resources and Evaluation, 2013, 47(1). - P.195-215.
5. Kneser R., Ney H. Improved backing-off for m-gram language modeling. // Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on (Vol. 1, P. 181-184). IEEE.
6. Shannon C. E. A Mathematical Theory of Communication // The Bell System Technical Journal, 1948, 27. - P. 379-423, 623-656.

Соловьев Федор Николаевич. Аспирант Московского политехнического университета. Окончил Московский физико-технический институт (государственный университет) в 2014 году. Область научных интересов: компьютерная лингвистика, автоматическая обработка текстов, распознавание образов. E-mail: the0@yandex.ru

Чеповский Андрей Михайлович. Профессор кафедры информационной безопасности Национального исследовательского университета «Высшая школа экономики» и кафедры прикладной математики и моделирования систем Московского политехнического университета. Доктор технических наук. Автор более 100 печатных работ. Область научных интересов: автоматическая обработка текста, информационный поиск, кибернетика. E-mail: achepovskiy@hse.ru

An extension of the short text language identification model

F.N Solovyev, A.M. Chepovskiy

Abstract: In our work we address the problem of the natural language identification in short texts. A Bayesian classifier is employed. We propose an extension of the language identification model by the incorporation of the new cyrillic languages of the russian small nations.

Keywords: statistical language model, natural language identification, languages of russian small nations.

References

1. Gusev S.V., Chepovskiy A.M. Natural language identification model // Business-informatics. 2011. No3 (17).
2. Zaidelman L.Y., Krylova I.V., Orekhov B.V. The technology of web-texts collection of Russian minor languages // In proceedings of International conference CPT2015, 2015. Moscow Region, Prorvino, ICPT, 2016. P. 179-181.
3. Vogel, J., Tresner-Kirsch, D. Robust language identification in short, noisy texts: Improvements to LIGA // Proceedings of the Third International Workshop on Mining Ubiquitous and Social Environments, 2012. - P. 43-50.
4. Carter, S., Weerkamp, W., Tsagkias, M. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text // Language Resources and Evaluation, 2013, 47(1). - P.195-215.
5. Kneser, R., Ney, H. Improved backing-off for m-gram language modeling. // Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on (Vol. 1, P. 181-184). IEEE.
6. Shannon C.E. A Mathematical Theory of Communication // The Bell System Technical Journal, 1948, 27. - P. 379-423, 623-656.

Solovyev Fedor Nikolayevich. Phd student at Moscow Polytechnical University, has master degree in Applied Maths and Informatics, graduated Moscow Physical Technical Institute (state university) in 2014, area of special scientific interest: computational linguistics, natural language processing, pattern recognition. E-mail: the0@yandex.ru

Chepovskiy Andrey Mikhailovich. Doctor of Tecnhical Science, professor of chair of information security National Research University «Higher School of Economics»; professor of chair of applied mathematics and modeling of systems at Moscow Polytechnical Universityauthor, author of 100 publications, area of special scientific interest: natural language processing, information retrieval, mathematical cybernetics, cybernetics. E-mail: achepovskiy@hse.ru