УДК 004.8

ОПРЕДЕЛЕНИЕ ЯЗЫКА ЭЛЕКТРОННОГО СООБЩЕНИЯ В РАМКАХ ЗАДАЧИ КОНТЕКСТНОГО АНАЛИЗА И КЛАССИФИКАЦИИ СПАМА

М. А. Артемов, В. А. Сорокина

Воронежский государственный университет

Поступила в редакцию 18.11.2011 г.

Аннотация. Рассматривается проблема распознавания естественного языка, на котором написано электронное письмо, в рамках проблемы борьбы со спамом и классификации электронных сообщений. Предложен новый способ определения языков на основе областей стандарта Юникод и статистических словарей.
Ключевые слова: спам, определение языка, контекстный анализ, классификация текста.
Annotation. In the paper the problem of language detection is analyzed as a part of a problem of email message classification and spam fight. The new method of language detection is suggested which is based on Unicode code pages and statistical dictionaries.
Key words: Spam, language detection, context analysis, text classification.

ВВЕДЕНИЕ

В рамках проблемы борьбы со спамом возникает задача классификации электронных писем, разбиение их по группам. Один из ключевых моментов в этой работе является определение естественного языка, на котором написано сообщение. Зная язык, возможно более точно определить группу к которой относится данное письмо. Кроме того, некоторые языки могут быть нежелательными и, следовательно, информация о языке будет достаточным признаком, чтобы отметить данное сообщение как спам и избежать проведения дополнительных контекстных проверок содержимого письма.

Предлагаемый метод определения языка сообщения основывается на использование областей стандарта кодирования символов Юникод и на применении статистических словарей двух типов, построенных особым образом.

Кодирование символов. Прежде всего, необходимо отметить, что информация о кодировке должна присутствовать внутри любого электронного письма. Это необходимо для правильного отображения текста электронного сообщения почтовым клиентом. Существует большое число различных кодировок для различных письменностей. Все они могут быть конвертированы в формат Юникод, представляющий все множество письменных символов используемых в разных языках. Если кодировка для некоторого письма не указана будем считать, оно использует кодировку UTF-8, которая является реализацией стандарта Юникод.

Коды символов стандарта Юникод, разделены на несколько областей [1]. В рамках задачи анализа текста интерес представляют только часть этих областей. А именно,

Latin, Cyrillic, Armenian, Greek, Coptic, Georgian, Ethiopic, Tifinagh, Arabic, Hebrew, Syriac, Cherokee, CanadAboriglSyl, Deseret, Shavian, Bengali, Devanagari, Gujarati, Gurmukhi, Kannada, Limbu, Malayalam, Oriya, Sinhala, Tamil, Telugu, Buhid, Tagalog, Tagbanwa, Buginese, Khmer, Lao, Myanmar, Thai, HanIdeograph, Hiragana, Katakana, Hangul, Yi, Mongolian.

Т.е. каждая из этих областей содержит символы определенной письменности.

На основе таблиц Юникода (http://www.unicode.org/charts/) было построено соответствие областей Юникод конкретным естественным языкам. (Для экономии места здесь приводится только часть соответствия «область Юникода — язык»).

Latin – ENGLISH ... FINNISH FRENCH ... GERMAN ...SOMALI SPANISH SWAHILI SWEDISH TAGALOG TAJIK TAMAZIGHT ...

Cyrillic – RUSSIAN ABKHAZ ... UKRAINIAN

Armenian – ARMENIAN

Greek – GREEK
Coptic – ARMENIAN
Georgian – GEORGIAN
Hebrew – LADINO HEBREW YIDDISH
Devanagari – BIHARI HINDI KASHMIRI KONKAN LIMBU MARATHI NEPALI SANSKRIT NEWARI
HanIdeograph – JAPANESE KOREAN CHINESE
Hangul – KOREAN
Hiragana – JAPANESE
Katakana – JAPANESE
...

Из приведенного выше соответствия видно, что для некоторых языков существует однозначная взаимосвязь: «область Юникода — язык». Например, это верно для греческого и грузинского языков. Однако такого соответствия нет для наиболее широко используемых латинских символов, так как все европейские языки, такие как французский, немецкий, испанский, а также некоторые экзотические языки используют латинский алфавит. Более того, некоторые языки, например, таджикский, используют сразу два алфавита: кириллический и латинский. Таким образом, зная, что некоторое слово написано латинскими, символами, нельзя однозначно определить, к какому именно языку его отнести. Для таких письменностей необходимо провести дополнительное исследование с помощью словарей.

Построение и использование статистических словарей. В зависимости от области Юникода каждое слово будет искаться в соответствующей группе словарей. Но здесь возникает следующая проблема: языки могут иметь совпадающие по написанию слова. Особенно это характерно для родственных языков, например, слово «трамвай» — общее для болгарского и русского, или «abnegado» — для испанского и итальянского. Для решения данной задачи введем два типа словарей. Словари первого типа содержат уникальные слова, однозначно определяющие язык. В словаре второго типа хранится информация о принадлежности данного слова нескольким языкам. А именно, пары: уникальный номер языка, в котором это слово встречается, и вес данного слова для этого языка, рассчитанный на основе статистических данных. Например, Admittance — 01 (английский), 43, 03 (французский), 56.

Для построения словарей пересечений (словарей второго типа) применялись архивы Википедии (http://static.wikipedia.org), содержащие большие объемы текстовой информации, категоризованной по языкам. Вес некоторого слова для каждого языка рассчитывался на базе статистической информации, извлеченной из текстов Википедии. Кроме того, очень популярные слова, такие как «in», встречающиеся в слишком большом количестве языков, не учитывались в формировании словарей.

Вынесение решения о принадлежности к языку. Собрав данные о возможной принадлежности слов письма к тому или иному языку, необходимо вынести решение о языке письма в целом. Для этого необходимо определить следующие величины: (T — степень участие i языка, зависящая от того в словаре какого типа найдено большинство слов для данного языка, L_i — вес i языка).

$$L_i = \begin{cases} 100 \cdot \left(Xpure_i + \left(1 + \dfrac{Xpure_i}{N} \cdot K_1 \right) \cdot Xovp_i \right), \\ \text{если } Xpure_i \neq 0, \\ 100 \cdot Xovp_i \cdot K_2, \text{если } Xpure_i = 0. \end{cases} \quad (1)$$

$$T = \sum_{i=1}^{n} \begin{cases} \left(1 + \dfrac{Xpure_i}{N} \cdot K_1 \right) \cdot Xovp_i; \\ \text{если } Xpure_i \neq 0, \\ Xovp_i \cdot K_2, \text{ если } Xpure_i = 0, \end{cases} \quad (2)$$

где – n общее количество определяемых языков, $Xpure_i$ — суммарный вес слов, найденных в словарях первого типа, для i языка, $Xovp_i$ – сумма весов для найденных слов из словаря второго типа, принадлежащих i языку, K_1 и — константы, с помощью которых можно влиять на степень участия уникальных и общих для нескольких языков слов в построении итоговой оценки. N — величина, рассчитанная следующим образом:

$$N = N_{pure} + T + N_{unknown} * K_{unknown}, \quad (3)$$

где N_{pure} — количество слов, найденных в словарях с уникальными словами, N_{ovp} — число слов, найденных в словаре пересечений, $N_{unknown}$ — количество «неизвестных» слов, которые не встретились ни в одном из словарей, $W_{unknown}$ — вес, с которым будут учитываться

«неизвестные» слова. С помощью $W_{unknown}$ можно снизить итоговую оценку для каждого языка, если процент «известных» слов получился небольшой.

Таким образом, итоговым языком письма считается i язык, набравший наибольший окончательный вес L_i, и этот вес превысил некоторый порог. В противном случае, язык письма неизвестен. Необходимо заметить, что в электронной рассылке достаточно распространенным является факт, когда в письме одновременно присутствует текст на двух языках, например английском и испанском или английском и русском. В таком случае можно также выделять группы-билингвы, в которых суммарный вес двух языков, имеющих максимальные оценки L_i, больше некоторого порога, причем сами эти оценки по отдельности этот порог не превышают.

Результаты тестирования. Рассмотрим результаты работы описанного алгоритма на общедоступном множестве писем, TREC CORPUS 06 (http://plg.uwaterloo.ca/~gvcormac/treccorpus06/). Таблицы 1 и 2 представляют собой результаты работы алгоритма на «спамной» и легальной частях корпуса писем. Это неполный вывод работы алгоритма. Для краткости здесь представлены результаты только для групп, состоящих более чем из 10 писем для спама и более чем из 30 – для не спама. Тест проходил в двух вариантах: со словарями обоих типов и без словарей (определение языка происходит только на основе Юникод областей, к которым принадлежат символы слов обрабатываемых сообщений). Язык «unknown» говорит о том, что данное письмо или содержит недостаточное количество слов (меньше пяти), или результат для каждого из определенных языков слишком низкий (меньше 50).

Из рассмотрения таблицы 1 следует, что большая часть «спамных» писем тестируемого множества написана на латинице и преобладающий язык английский. Таблица 2 представляет распределение легальной части корпуса. Большая часть писем, как и в спаме, написана на латинице, но также здесь присутствует значительная часть писем (15%), которые используют японские (katakana) символы.

Таблица 1

Результаты определения языков для писем спам-части TREC CORPUS 06

Со словарями двух типов			Без словарей (на основе областей Юникода)		
English	12649	(97.9783 %)	latinX	12818	(99.2874 %)
Unknown	106	(0.8210 %)	Unknown	85	(0.6584 %)
english_spanish	25	(0.1936 %)	hangulX	4	(0.0309 %)
english_french	19	(0.1471 %)	katakanaX	1	(0.0077 %)
French	11	(0.0852 %)	cyrillicX	1	(0.0077 %)
english_italian	11	(0.0852 %)	bengali_latinX	1	(0.0077 %)
Czech	11	(0.0852 %)			

Таблица 2

Результаты определения языков для писем легальной части TREC CORPUS 06

Со словарями двух типов			Без словарей (на основе областей Юникода)		
English	17263	(69.2959 %)	latinX	18115	(72.716 %)
Japanese	4113	(16.5101 %)	katakanaX	3794	(15,070 %)
Unknown	1395	(5.5997 %)	Unknown	1532	(6.1497 %)
Chinese	826	(3.3157 %)	hanideographX	803	(3.2234 %)
Russian	558	(2.2399 %)	cyrillicX	563	(2.2599 %)
Portuguese	170	(0.6824 %)	hangulX	56	(0.2271 %)
Korean	52	(0.2087 %)	thaiX	36	(0.2271 %)
english_french	45	(0.1806 %)	hebrewX	12	(0.0482 %)
French	41	(0.1646 %)	devanagariX	1	(0.0040 %)
english_irish	38	(0.1525 %)			
Thai	34	(0.1365%)			
chinese_english	32	(0.1285 %)			

ЗАКЛЮЧЕНИЕ

Предложенный метод показывает высокую точность результата, что, конечно, является большим преимуществом данного алгоритма. Однако для его работы необходимы словари, которые могут занимать большее количество места на диске. Так, для описанного теста были построены словари, занимающие порядка 100 мб. Вероятным продолжением данной работы может быть исследование возможности построения словарей с меньшим объемом данных или работа только с одним из типов словарей.

СПИСОК ЛИТЕРАТУРЫ

1. The Unicode Consortium. The Unicode Standard, Version 6.0.0 // Mountain View, CA: The Unicode Consortium, 2011.

Артемов Михаил Анатольевич – заведующий кафедрой математического обеспечения и администрирования информационных систем Воронежского государственного университета, доктор физико-математических наук, профессор. E-mail: atremov_m_a@mail.ru

Сорокина Виктория Александровна – аспирант кафедры «Программное обеспечение и администрирование информационных систем», факультет «Прикладная математика, информатика и механика», Воронежский Государственный Университет. Email: vica.sorokina@gmail.com

Artemov Mikhail A. – Head of Department Software & Information System Administering, Voronezh State University, doctor of Physics-math. Sciences, Professor. E-mail: atremov_m_a@mail.ru

Sorokina Viktoriya A. – Postgraduate student, the department of "Software and information system administering", Voronezh State University. Email: vica.sorokina@gmail.com