МОДЕЛЬ ДЛЯ ИДЕНТИФИКАЦИИ ЕСТЕСТВЕННОГО ЯЗЫКА ТЕКСТА

С.В. Гусев,
программист ЗАО «НОРСИ-ТРАНС»

А.М. Чеповский,
кандидат технических наук, профессор кафедры информационных систем Московского государственного университета печати (МГУП);
доцент кафедры анализа данных и искусственного интеллекта НИУ ВШЭ

Адрес: г. Москва, ул. Прянишникова, д. 2а
E-mail: unk379@mail.ru, achep@adde.math.msu.su

В статье рассмотрена проблема автоматической идентификации естественного языка текста. Предлагается статистическая модель текстов на естественном языке. Рассматриваются алгоритмы определения естественного языка текста.

Ключевые слова: статистическая модель языка, идентификация естественного языка текста.

Введение

В информационных системах различного типа, предназначенных для обработки в автоматическом режиме больших объемов текстов на естественных языках, актуальны различные задачи распознавания текстовой информации. Требование автоматизации процессов обработки текстовой информации придают особую важность проблемам определения естественного языка, на котором написан текст, или часть текста.

Для информационных систем предприятий, систем документооборота особую важность приобретает задача определения в автоматическом режиме языка коротких текстовых сообщений размеров, характерных для сообщений электронной почты и аналогичных коммуникационных сервисов. В настоящее время известны достаточно точные методы распознавания языка для длинных текстов, содержащих десятки слов и предложений [1]. Модели, использующие частоты буквосочетаний, широко использовались для определения языка текста [2, 3]. В [3] отмечалось, что возможно использовать ранговые методы для идентификации языка текста,

но они не применимы для коротких текстов. В работе [3] делается вывод о том, что проблема определения языка коротких сегментов текста остается актуальной, а достижение более высокой точности осуществлялось за счет более крупных моделей и медленной скорости классификации.

В данной статье рассматривается статистическая модель строки текста на естественном языке и разрабатывается методика ее реализации с целью достижения эффективности использования статистики буквосочетаний символов языка для достижения высокой точности классификации текстов по используемым естественным языкам.

Статистическая модель текста на естественном языке

Задача определения языка текста является задачей распознавания образов и может решаться на базе вероятностной модели. Принцип Байесовского классификатора можно применить к строке символов, считая, что нам известны статистические характеристики для символов в текстах на конкретном естественном языке, или текстах, относящихся к заданному классу.

Рассмотрим строку s, состоящую из N символов c_n (n = 1, ..., N), принадлежащих алфавиту Σ. Конкретное значение строки будем обозначать следующим образом: $s = <c_1\ c_2\ ...\ c_N>$, а значение отдельного символа строки на i-той позиции будем обозначать так: $s[i] = c_i$. Для решения задачи необходимо отнести данную строку к одному из классов Y_l (l = 1, ..., K), где под Y_l (l = 1, ..., K) понимается один из K языков.

Считаем, что каждый класс задает некоторое распределение вероятностей на множестве всех допустимых строк. В таком случае возможно применение статистического критерия максимального правдоподобия для определения класса, которому принадлежит классифицируемая строка.

Вероятность появления строки s в некотором языке равна произведению вероятностей появления каждого из её символов при условии появления всех символов, идущих перед ним:

$$\begin{aligned} &P(s = c_1...c_N) = P(s[N] = c_N \mid s[1] = \\ &= c_1, ..., s[N-1] = c_{N-1}) * \\ &P(s[N-1] = c_{N-1} \mid s[1] = c_1, ..., s[N-2] = \\ &= c_{N-2}) * ... * P(s[1] = c_1) \end{aligned} \quad (1)$$

Предположим, что распределение вероятностей i-того символа, зависит не более чем от k предыдущих, тогда формула (1) примет вид:

$$\begin{aligned} &P(s[i] = c_i \mid s[1] = c_1, ..., s[i-1] = c_{i-1}) = \\ &= P(s[i] = c_i \mid s[i-k] = c_{i-k}, ..., s[i-1] = c_{i-1} \end{aligned} \quad (2)$$

По обучающей выборке производится оценка условных вероятностей. Для этого вычисляются частоты всех подстрок длины не более k, и значения условных вероятностей появления очередного символа принимаются равными отношению частот соответствующих подстрок:

$$\begin{aligned} &P(Y_l, s[i] = c_i \mid s[i-m] = c_{i-m}, ..., s[i-1] = \\ &= c_{i-1}) = \frac{f(c_{i-m}...c_i)}{f(c_{i-m}...c_{i-1})},\ \forall m \le k \end{aligned} \quad (3)$$

где $f(c_i)$ – частота встречаемости подстрок в обучающей выборке.

Для каждой классифицируемой строки вычисляется ее вероятностная оценка относительно каждого класса:

$$\begin{aligned} &P(Y_l, s) = P(Y_l, s[N] = c_N \mid s[N-k] = \\ &= c_{N-k}, ..., s[N-1] = c_{N-1}) * \\ &P(Y_l, s[N-1] = c_N \mid s[N-k-1] = \\ &c_{n-k-1}, ..., s[N-2] = c_{N-2}) * ... * P(Y_l, s[1] = c_l) \end{aligned} \quad (4)$$

Классифицируемая строка относится к классу, относительно которого она имеет наибольшую оценку.

Алгоритм реализации модели для идентификации языка

Каждый текст на естественном языке подвергается предварительной обработке, в результате которой из текста получается набор слов, состоящих только из символов алфавита соответствующего языка, приведённых к нижнему регистру. По полученному набору слов формируется частотный словарь буквосочетаний длиной от 1 до n с учётом количества вхождений слов в частотный словарь. Данная процедура выполняется при построении модели строки конкретного языка, которая строится на основе массива обучающих текстов и завершается заполнением базы данных профилей для каждого исследуемого языка.

К каждому слову текста прибавляется один пробел справа, и оно подаётся на вход автомата. Об-

работка слова начинается из начального состояния автомата, соответствующего буквосочетанию, состоящему из (*n* - *1*) пробела. Для очередного символа вычисляется вероятность перехода по нему из текущего состояния в следующее состояние по формулам (1) – (4). Вероятностью появления данного слова считается произведение вероятностей всех переходов, произошедших во время обработки слова автоматом.

Для определения языка текста оценивается вероятность соответствия рассматриваемого текста моделям строк для каждого естественного языка. Выбирается максимальная вероятность, которая соответствует языку, на котором написан текст.

Предложенная методика и алгоритмы реализованы в виде библиотеки для платформы MS Windows x86.

Точность определения языка текста

Для фрагмента текста может быть вычислена числовая оценка того, насколько данный фрагмент соответствует модели строки текста на естественном языке. Пусть фрагмент текста *s* содержит *N* символов. Вероятность его появления в тексте на *l*-том языке может быть оценена в соответствии с формулой (4). Тогда оценку соответствия этого фрагмента *l*-тому языку будем определять следующим образом:

$$E_l(s) = \frac{\ln(P(Y_l, s))}{N} + const \qquad (5)$$

где $P(Y_l, s)$ – вероятность появления строки *s* в языке Y_l;

N – длина строки *s* в символах;

const – нормирующая постоянная величина.

Математическое ожидание такой оценки не зависит от длины фрагмента текста.

При вычислении оценки соответствия данному языку фрагментов текстов на других языках результат сильно зависит от используемого другими языками алфавита и степени похожести языков на данный язык. На *рисунке 1* представлены распределения оценок (5) соответствия модели строки на русском языке для 5000 фрагментов длины 20 символов из текстов на русском, белорусском и английском языках.

Видно, 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 (белорусский), то невозможно точно отделить тексты на другом языке при простейшем пороговом отсечении по максимальному значению вероятности.

При увеличении длины фрагмента шансы отделения неизвестного языка с помощью порогового значения увеличиваются. На *рисунке 2* показано распределение оценок 2000 фрагментов длиной 200 символов для текстов на русском и белорусском языках. Видно, что для длины фрагмента текста в 200 символов уже можно отделить русский текст от белорусского с высокой степенью достоверности.

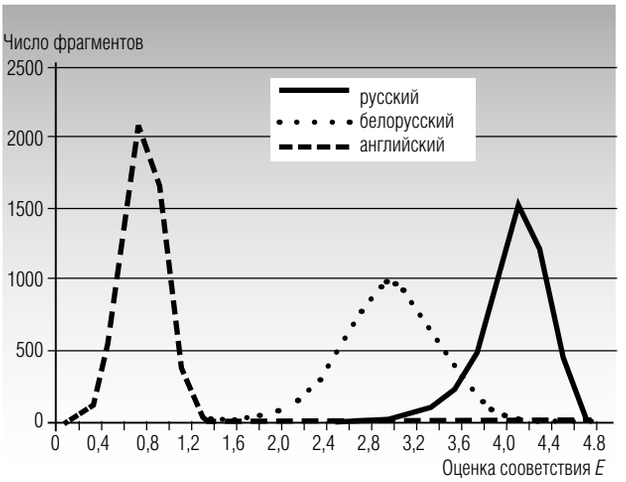

Рис. 1. Распределение оценки соответствия модели текста по количеству фрагментов длиной 20 символов для различных языков.

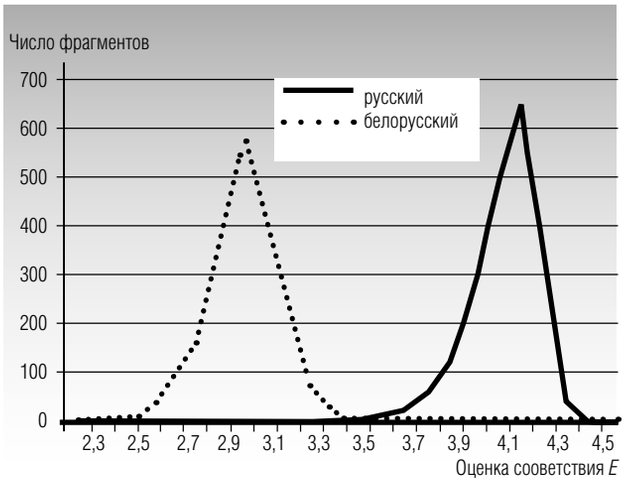

Рис. 2. Распределение оценки соответствия модели текста по количеству фрагментов длиной 200 символов для различных языков.

Приведенные результаты на *рисунках 1* и *2* показывают существование серьезной проблемы при определении языков в условиях, когда входные тексты могут содержать языки, отсутствующие среди определяемых языков.

Для отсечения неизвестных языков используется пороговое значение, которое вычисляется отдельно для каждого языка и каждой длины фрагмента:

$$T_l = M[E_l] - \gamma * \sigma[E_l], \qquad (6)$$

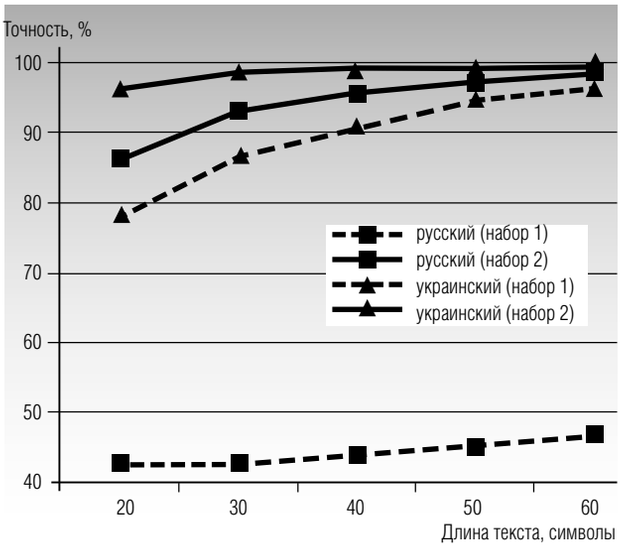

Рис. 3. Зависимость точности определения русского и украинского языков от длины текста для двух вариантов обучающих наборов.

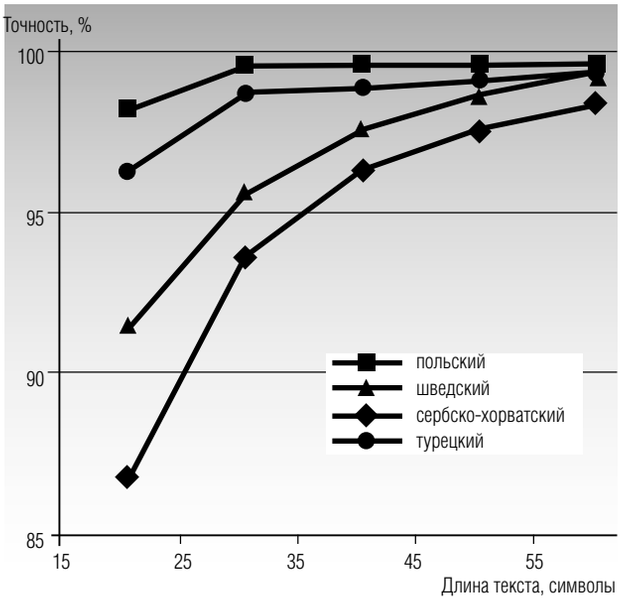

Рис. 4. Зависимость точности определения языка в зависимости от длины текста.

где $M[E_l]$ – математическое ожидание оценки соответствия для *l*-того языка;

$\sigma[E_l]$ – среднеквадратичное отклонение оценки соответствия для *l*-того языка;

γ – настраиваемый коэффициент порога отсечения.

Для высокой точности определения языка в условиях, когда входные тексты могут содержать неизвестные языки, требуется уметь определять все языки, использующие такую же систему письменности (т.е. алфавиты должны существенно пересекаться). Тексты, написанные на языках, использующих другие системы письменности, можно отсечь с помощью порогового значения оценки даже в случае очень коротких фрагментов. Для увеличения качества распознавания языков необходимо использовать базы данных для наибольшего количества языков.

Экспериментальные исследования

Для оценки точности распознавания используются следующие характеристики. Коэффициент релевантности определяет, какова в результирующем наборе, принадлежащем определенному классу, доля файлов, которые действительно являются файлами на соответствующем естественном языке. Коэффициент полноты измеряет при тестовом эксперименте, какая доля файлов данного типа (на соответствующем естественном языке) из тестового набора правильно отнесена к данному типу файлов. В качестве главной характеристики качества распознавания используется усредненная точность, которая определяется как взвешенное гармоническое среднее коэффициента релевантности и коэффициента полноты.

Всего рассматривалось сорок пять естественных языков, включающих использующие кириллическую письменность языки (славянские и относящиеся к языкам тюркской и иранской групп), основные языки индоевропейской семьи, использующие латинскую письменность. Эксперименты по определению естественного языка текста проводились на тестовых массивах, суммарные размеры которых в зависимости от языка изменялись от 1Мб до 7Мб «чистого» текста. При этом 80% текстов использовалось для обучения, а 20% этих корпусов текстов рассматривались как тестовые наборы, язык которых распознавался.

Влияние обучающих выборок на качество распознавания языков изучалась на двух наборах естественных языков, использующих кириллическую письменность. Набор 1 состоял из шести славянских языков: Русский, Украинский, Белорусский, Болгарский, Сербскохорватский, Македонский. Набор 2 включал языки первого набора и десять других языков, использующих кириллическую письменность: Азербайджанский, Башкирский, Карачаево-балкарский, Казахский, Киргизский, Татарский, Узбекский, Осетинский, Таджикский, Монгольский. Идентификация языка текста проводилась для двух вариантов обучения, основанных на описанных выше двух наборах естественных языков.

На *рисунке 3* представлены зависимости усредненной точности определения русского и украинского языков в зависимости от размера анализируемых текстов, измеряемых в количестве алфавитных символов языка, исключая пробелы. Видно, что при увеличении обучающего набора с «Набор 1», включающего 6 языков, до «Набор 2», включающего 16 языков, точность определения языков существенно увеличивается.

Тестовые эксперименты по исследованию качества распознавания естественных языков проводились для 45 естественных языков. Для всего этого набора проводилось обучение системы распознавания языков, и идентифицировались языки для текстов на всех 45 языках. Типичные результаты представлены на *рисунке 4*, где приведены результаты численных экспериментов на примерах распознавания польского, сербскохорватского, шведского и турецкого языков в зависимости от размеров анализируемых текстов. Достигнутые показатели качества аналогичны представленным в [3]. Видно, что для текстов даже небольших размеров достигается высокая (близкая к 100%) точность определения естественного языка.

9. Заключение

Таким образом, в статье предложена вероятностная модель строки текста на естественном языке и разработана методика автоматического определения языка текстового сообщения на естественном языке, включая небольшие по объему тексты.

Разработанная модель текста на естественном языке и методика ее реализации показала эффективность использования статистики буквосочетаний символов языка для решения задач распознавания текстов на естественном языке. Достигнута высокая точность распознавания естественного языка текста.

Методика эффективна при обработке больших объемов текстовой информации на различных естественных языках без использования словарей и грамматик этих языков. Предложенная методика может быть использована при решении практических задач автоматической обработки текстовой информации в информационных системах предприятий и организаций. ■

Литература

1. Paul B., McNamee. Language identification: a solved problem suitable for undergraduate instruction // Journal of Computing Sciences in Colleges, 2005, 20(3). – P. 94-101.
2. Cavnar W.B., Trenkle J.M. N-gram-based text categorization // Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. – Las Vegas, US, 1994. – P. 161-175.
3. Vatanen T., Väyrynen J., Virpioja S. Language identification of short text segments with n-gram models // Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), 2010. – P. 3423–3430.