Раздел III. Защита телекоммуникаций

УДК 681.056

А.Ю. Кулай, Д.А. Леднов, С.Ю. Мельников

О СТАТИСТИЧЕСКИХ МЕТОДАХ ИДЕНТИФИКАЦИИ ЯЗЫКА ИСКАЖЕННЫХ ТЕКСТОВЫХ И РЕЧЕВЫХ СООБЩЕНИЙ

Рассмотрены различные статистические методы идентификации языка искаженных текстов. Приведено экспериментальное сравнение их эффективности для различных длин текстовых сообщений. В предположении, что предложенная модель искажений последовательности знаков текста адекватна искажениям, наблюдаемым при обработке речевого сигнала, приводятся рекомендации по выбору статистических методов в задаче идентификации языка речевого сообщения.

1. Введение

В ряде систем автоматической обработки речи в качестве входного блока обработки речевого сигнала используются так называемые фонетические распознаватели, которые преобразуют последовательность отсчетов речевого сигнала (например, в формате 16 бит, 8КГц) в символьную последовательность в алфавите фонем того или иного языка [1]. Как правило, такие распознаватели работают со значительными ошибками. Ряд авторов для моделирования выходных последовательностей фонетических распознавателей предлагает использовать искаженные случайным образом последовательности букв текста. Такое моделирование, конечно, не вполне точно отражает реальную ситуацию, но значительно упрощает работу исследователя, исключая трудоемкий (связанный с наличием размеченных речевых корпусов) этап построения фонетических распознавателей.

Идентификация языка речевых и искаженных текстовых сообщений является частным случаем задачи распознавания образов, для решения которой требуется построить статистический критерий принадлежности нового сообщения к одному из классов, задаваемых «обучающими» сообщениями.

Выделяют две задачи идентификации языка: закрытую и открытую. В закрытой задаче идентификации предполагается, что новое сообщение принадлежит одному из рассматриваемых языков, в открытой – новое сообщение может быть на неизвестном языке. Задачу идентификации можно рассматривать как построение статистического критерия для конечного числа простых гипотез в случае закрытой задачи или для конечного числа простых и одной сложной гипотезы (сообщение на неизвестном языке) в случае открытой задачи.

При решении задачи идентификации языка и ряда других прикладных задач для описания статистических свойств последовательностей обычно используются *s*-граммные модели небольших порядков [2] в сочетании с методами сглаживания вероятностей [3], в которых при вычислении вероятностей *s*-грамм старших порядков используются частоты встречаемости в обучающем множестве *s*-грамм меньших порядков.

В работе [4] рассматривается оригинальный метод идентификации языка на искаженных разметках речевой базы, при котором используются словари (pattern tables) как в алгоритме сжатия данных LZW [5].

Одним из перспективных и активно исследуемых в последнее время инструментов, применяемых для ряда задач распознавания образов, является метод «двоичных деревьев» (binary tree, BT) [6]. Этот метод основан на уменьшении сложно-

сти модели за счет кластеризации контекста. Вероятность текущего наблюдения обуславливается множеством (кластером) предыстории символа. Кластеры, вообще говоря, могут включать предыстории различной длины, поэтому может учитываться информация из контекстов большой длины при относительно небольшой сложности модели.

Для некоторых специфических случаев задачи идентификации языка, например, при анализе результатов работы систем автоматической обработки текстов, могут быть применены методы теории автоматов [7]. Однако такие методы характеризуются весьма значительной вычислительной сложностью.

2. Идентификация языка искаженного текста

Рассматривалась закрытая задача идентификации английского, испанского, польского и французского языков. Эксперименты проводились с текстами рассказов А. Конан-Дойля о Шерлоке Холмсе, записанными в латинице без пробелов и знаков препинания в одном регистре, таким образом, алфавит состоит из 26 символов. Предложения рассматривались отдельно. Тексты разбиты на три подмножества: train set (английский – $2{,}3 \cdot 10^6$ символов, испанский – $1{,}8 \cdot 10^6$ символов, польский – $0{,}3 \cdot 10^6$ символов, французский – $0{,}8 \cdot 10^6$ символов), development set (по $50 \cdot 10^3$ символов) и test set (по $24 \cdot 10^3$ символов). На train set обучались модели языков, на development set вычислялись дополнительные параметры, на test set проводились эксперименты.

Текст подвергался следующим искажениям: пропуск, вставка и замена. С вероятностью 0,15 происходит пропуск символа. С вероятностью 0,15 происходит вставка символа, выбор которого осуществляется случайно равновероятно. С вероятностью 0,3 происходит равновероятная замена символа латинского алфавита на любой другой. Данные искажения можно считать близкими к тем ошибкам, которые допускают реальные фонетические распознаватели [8].

3. Методы идентификации

3.1. Back-off

На обучающем множестве строятся s-граммные модели (s = 2, 3, 4) с применением back-off метода сглаживания вероятностей [3]. На выходе языковых моделей получается 12-мерный (по 3 модели для 4 языков) вектор вероятностей, который нормируется и затем поступает на вход гауссовского конечного классификатора (Gaussian back-end classifier) [2]. Конечный классификатор содержит четыре (по одному для каждого языка) 12-мерных нормальных распределения с диагональной ковариационной матрицей. Параметры распределений (математические ожидания и дисперсии) оцениваются на development set. Для поступившего на вход конечного классификатора 12-мерного вектора для каждого языка вычисляется плотность распределения в соответствующей точке. За истинный принимается язык с максимальным значением плотности распределения.

3.2. PPM

Строятся s-граммные модели (s = 2, 3, 4) с применением сглаживания вероятностей по методу «C» алгоритма сжатия PPM [9]. Метод «C» является одним из наиболее часто используемых при сжатии данных [10]. Существенная разница с back-off методом Катца заключается в том, с каким весом для вычисления вероятностей s-грамм старших порядков берутся частоты встречаемости s-грамм меньших порядков. На выходе языковых моделей получается 12-мерный (по 3 модели для 4 языков) вектор вероятностей, который нормируется и затем поступает на вход конечного классификатора, как и в случае back-off.

3.3. LZW

На обучающем множестве строятся зависимые от языка словари (с максимальной длиной «слова» – 10 символов), т.е. множества встречающихся последовательностей длиной до 10 символов. Для тестируемого предложения вычисляется степень сжатия CR (compression ratio) или статистика WDS (weighted discriminant score), на основе которых принимается решение о языке [4].

3.4. Двоичные деревья

На обучающем множестве для каждого языка строится двоичное дерево с 3 предикторами. Параметры дерева оптимизируются на development set [10]. Для тестируемого предложения вычисляется вероятность в соответствии с каждым двоичным деревом. За истинный принимается язык с максимальным значением вероятности.

3.4.1. Описание и построение двоичного дерева

Двоичное дерево состоит из множества неконечных и конечных вершин. Каждая неконечная вершина ассоциируется с двоичным тестом (вопросом) и имеет два перехода в вершины следующего яруса. Каждая конечная вершина (лист) помечена распределением на алфавите символов. Для того чтобы посчитать вероятность символа a_t в момент времени t с предысторией $a_{t-1},...,a_{t-N}$, необходимо пройти из корня по неконечным вершинам графа по пути, определяемым ответами на двоичные тесты, пока не встретится конечная вершина. Вероятность символа a_t получается из распределения, которым помечен этот лист. Пример двоичного вопроса: «$a_{t-3} \in \{[A],[E],[I],[O],[U],[Y]\}$?». Итак, путь по графу определяется предысторией текущего символа.

Опишем построение двоичного дерева. Пусть есть обучающее множество $A = \{a_1,...,a_T\}$ с распределением $Y_A = \{p(a_j / A)\}_{1 \le j \le K}$, где $a_j \in \mathrm{A}$ – алфавит символов мощности K. Основным шагом при построении двоичного дерева является разбиение на два подмножества $A_1 \cup A_2 = A$, используя которые в дереве создаются две вершины следующего уровня. Для оценки качества разбиения используется энтропия

$$H(Y_A) = -\sum_{j=1}^{K} p(a_j / A) \log_2 p(a_j / A).$$

Разбиение базируется на множестве «предикторов» (предыстории) (для a_t – это $a_{t-1}, a_{t-2},...$) каждого элемента из A и двоичном вопросе Q. Q может быть составным, но на практике обычно это выражение типа «$X \in S$?», где X – выбранный предиктор, например $X = a_{t-2}$, $S \subset \mathrm{A}$ – некоторое подмножество символов. Критерий разбиения ищет тот Q^*, при котором уменьшение средней энтропии максимально.

Рекурсивный алгоритм построения дерева (количество предикторов равно N):

1. Пусть n – текущая вершина. Изначально, n – корень.

2. Для каждого предиктора X_i ($i = 1,...,N$) ищем подмножество символов $S_i^n \subset A$, т.е. вопрос «$Q_i : X_i \in S_i^n$?», который минимизирует среднюю условную энтропию распределения звуков Y в вершине n:

$$H_i(Y) = p(Q_i)H(Y/Q_i) + p(\overline{Q_i})H(Y/\overline{Q_i}).$$

3. Определяем, какой из вопросов шага 2 дает наименьшую энтропию. Пусть это будет вопрос k, т.е.

$$k = \arg\min_{1 \le i \le N} H_i(Y).$$

4. Уменьшение энтропии в вершине n за счет вопроса k равно

$$H(Y) - H_k(Y).$$

Если уменьшение «существенно», то запоминаем вопрос k, создаем две вершины-потомка n_1 и n_2, пропускаем данные соответственно условиям $X_k \in S_k^n$ и $X_k \notin S_k^n$, и повторяем шаги 2-4 для новых вершин отдельно.

Для ускорения поиска подмножества S_i^n на втором шаге алгоритма используется вариант FF1 алгоритма Flip-Flop, предложенный в [12]. Данный алгоритм, как показано в [13], является достаточно быстрым и эффективным при решении задачи распознавания языков.

3.5. Многогранники автоматов

Укажем на принципиальную возможность использования методов теории автоматов для одной специфической задачи идентификации языка. Предположим, что текст на исходном языке обработан одной из систем автоматической обработки текста (например, с использованием автоматического перевода) A_1, A_2, …, A_t, и требуется определить, какой именно. Рассматривая A_1, A_2, …, A_t как конечные автоматы, задачу идентификации языка в такой постановке можно трактовать как задачу идентификации автомата с неизвестным входом. Для ее решения можно воспользоваться подходом, предложенным в [7].

Идея подхода состоит в следующем. Предположим, что имеется текст $\gamma^{(N)} = (y_0, y_1, ..., y_{N-1})$, и нужно определить, каким из автоматов A_1, A_2, …, A_t он порожден. Зафиксируем набор мультиграмм $\{\beta_j, j = 1,2,...,k\}$ в объединении выходных алфавитов автоматов. Пусть $p_{\beta_1}^{(N)},..., p_{\beta_k}^{(N)}$ – относительные частоты встречаемости мультиграмм $\beta_1,...,\beta_k$ в последовательности $\gamma^{(N)}$. В [7] приведен алгоритм, который автомату A_i ставит в соответствие определенный выпуклый многогранник R_{A_i} в кубе $[0,1]^k$. Доказано, что для R_{A_i} выполняется неравенство

$$\rho\left(\left(p_{\beta_1}^{(N)},\ldots,p_{\beta_k}^{(N)}\right),R_{A_i}\right)\le\frac{D_i+2(l-1)}{N+D_i+l-1},$$

где $l=\max\{|\beta_j|\}$, под расстоянием $\rho(u,v)$ между двумя точками $u,v\in R^k$ понимается максимум модулей разностей координат этих точек. D_i – диаметр графа переходов автомата A_i. Если это неравенство нарушается, то гипотеза о том, что текст был обработан системой A_i, бракуется.

Преимущества указанного метода в том, что не используется никаких предположений о вероятностной природе входного текста, в том числе предположения о его стационарности. К недостаткам этого метода следует отнести значительную вычислительную сложность построения многогранника.

4. Результаты идентификации языка искаженных текстов

В табл. 1 приводятся характеристики точности идентификации языка искаженных текстов для различных методов.

Таблица 1

Средний процент правильной идентификации языка искаженных текстов для различных методов

Длина тестируемого предложения в символах	Back-off	PPM	LZW	Двоичные деревья
1-20	56,3	49,59	41,49	48,55
21-40	71,32	69,91	55,8	71,32
41-60	79,16	75,99	59,29	80,58
61-80	83,33	80,54	67,57	85,95
81-100	90,57	86,55	74,37	93,28
более 100	93,32	93,21	68,11	94,65

5. Идентификация языка речевого сообщения

Рассматривалась закрытая задача идентификации восьми языков (английский, испанский, китайский, польский, русский, французский, хинди и японский). Система строилась по так называемой схеме PPRLM (рис. 1), с параллельным использованием нескольких фонетических распознавателей [2].

Для выходных последовательностей распознанных фонов строились 3- и 4-граммные модели с применением back-off метода сглаживания вероятностей. Решение о языке принимал гауссовский конечный классификатор. Для длительности входного сообщения 20 секунд средняя точность составила 84,21%, для длительности входного сообщения 40 секунд средняя точность составила 91,3%. С учетом примерной скорости фонемообразования 3-5 фонем в секунду такие данные попадают в диапазон табл. 1 и в целом согласуются с предложенной в п. 2 моделью искажений.

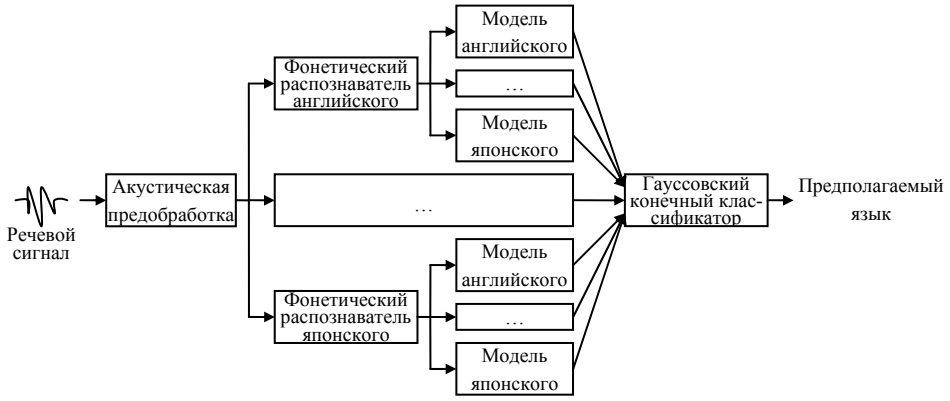

Рис. 1. Схема PPRLM

6. Заключение

Сравниваются четыре статистических метода идентификации языка искаженных текстовых сообщений: с использованием *s*-граммных моделей с двумя различными способами сглаживания вероятностей, на основе алгоритма сжатия данных LZW, двоичных деревьев. В проведенных экспериментах с искаженными текстами лучшие результаты продемонстрировал метод двоичных деревьев, метод на основе алгоритма сжатия данных LZW оказался хуже других. По результатам экспериментов на искаженных текстах можно рекомендовать метод двоичных деревьев для повышения точности идентификации языка речевых сообщений.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Батальщиков А.А., Леднов Д.А.* Модель открытой идентификации языка // Сб. трудов XVII сессии Российского Акустического Общества, 11-17 сентября 2006 г. Таганрог. – Москва, ГЕОС, 2006. Т. 3. – С. 44-45.

2. *Campbell W., Gleason T., Navratil J., Reynolds D., Shen W., Singer E., Torres-Carrasquillo P.* Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation. In Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, (San Juan, Puerto Rico), June 2006.

3. *Katz S.M.* Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech and Signal Processing, 35(3): 400-401, 1987.

4. *Basavaraja S.V., Screenivas T.V.* Low Complexity LID using Pruned Pattern Tables of LZW. In INTERSPEECH-2006, paper 1398-Mon2CaP.4.

5. *Nelson M.* LZW Data Compression. Dr Dobbs Journal, Oct 1989.

6. *Bahl L., Brown P., DeSouza P., Mercer R.* A tree-based statistical language model for natural language speech recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing, 37(7): 1001-1008, July 1989.

7. *Мельников С.Ю.* Многогранники, характеризующие статистические свойства конечных автоматов // Труды по дискретной математике, 2003. Т. 7. – М.: Изд-во физико-математической литературы, 2003. – С. 126-137.

8. *Kulay A.Y., Melnikov S.Y.* Different approaches to the garbled text language recognition, using the data compression methods. Proc. XII intern. Conference “Speech and Computer” 15-18 Oct. 2007, vol. 2, pp. 697-701.

9. *Moffat A.* Implementing the PPM data compression scheme. IEEE Transactions on Communications, 38(11): 1917-1921, 1990.

10. *Ватолин Д., Ратушняк А., Смирнов М., Юкин В.* Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. – М.: ДИАЛОГ-МИФИ, 2002.

11. *Кулай А.Ю., Мельников С.Ю.* Сравнение нескольких подходов к распознаванию языков искаженных текстов // Труды второй международной конференции «Системный анализ и информационные технологии» (САИТ-2007), (Обнинск, Россия), 10-14 сентября 2007 г. – М.: Изд-во ЛКИ, 2007. Т. 1. – С. 218-220.

12. *Nadas A., Nahamoo D., Picheny M., Powell J.* An iterative «Flip-Flop» approximation of the most informative split in the construction of decision tree. Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1991), (Toronto, Canada), May 1991, pp. 565-568.

13. *Navratil J.* Recent advances in phonotactic language recognition using binary-decision trees. In INTERSPEECH-2006, paper 1338-Mon2CaP.6.

УДК 681.327.8

Д.Ф.Хисамов

МОДЕЛИРОВАНИЕ СИНХРОНИЗАЦИИ ПСЕВДОСЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА КАНАЛАХ СВЯЗИ С ЗАВИСИМЫМИ ОШИБКАМИ

Постановка задачи

Пусть по каналу с аддитивной помехой передается рекуррентный сигнал (РС) длительностью в N символов. Прием РС осуществляется по “зачетному отрезку” [1]. Определим вероятность правильной синхронизации РС при наличии зависимых ошибок в канале. Рекуррентный сигнал на интервале анализа N’ можно разбить на блоки из ε элементов, кратных длине “зачетного отрезка” n , то есть

$$\varepsilon = n/J \quad , \quad J=1,2,...,n, \tag{1}$$

где J – параметр, указывающий, на сколько частей разбит “зачетный отрезок”.

Таких блоков на длине N’ может быть Z=N’•J/n.

Условимся блок называть непораженным, если все ε элементов блока приняты безошибочно, и пораженным при наличии хотя бы одной ошибки в блоке, и обозначим состояния блоков соответственно через 0 и 1. Тогда блочное отображение принимаемого РС можно представить двоичной последовательностью:

$$\overline{S} = S_1, S_2, \ldots, S_Z , \tag{2}$$

где:

$$S_i = \begin{cases} 0 \text{ } \textit{если блок непаражен}, \\ 1 \quad \textit{если блок поражен}, \end{cases}$$

а вероятности правильного приема РС будет соответствовать вероятность появления в последовательности $\overline{S}$ серии из J нулей подряд.

Допустим, что последовательность $\overline{S}$ апроксимируется односвязной цепью Маркова [2]. Чем больше длительность блока, тем эта аппроксимация будет точнее, так как при этом уменьшается зависимость между блоками, отстоящими друг