УДК 004.512

Система контроля достоверности текстовой информации на основе *n*-граммных парсинговых моделей

М. М. Камилов, А. Р. Ахатов

Ñàìàðêàíäñêèé ãîñóäàðñòâåííûé óíèâåðñèòåò èì. À. Íàâîè, 140104, Ñàìàðêàíä, Óçáåêèñòàí

Предложен новый подход к построению компьютерной системы передачи и обработки текстовой информа- ции на основе *n*-граммной языковой модели. Получены методики определения условных вероятностей *n*-кратных ошибок в информации, разработаны способы и алгоритмы оптимизации основных компонент системы контроля и коррекции орфографии, построенных на основе механизмов парсингового представле- ния и моделирования элементов текста.

The author offers a new approach for construction computer system of transfer and processing the text informa- tion on a basis of *n-*gram language model. In article it is stated results of receiving the techniques of definition con- ditional probabilities of *n-*multiple mistakes in the information. It is developed the ways and algorithms of optimi- zation the basic component of the monitoring system and spelling correction constructed on the basis of parsing representation mechanisms and modeling text elements.

1. **Постановка задачи контроля и коррекции текстовой информации.** Функционирование любых информа- ционных систем в существенной степени зависит от достоверности передачи сообщений, которая снижается вследствие ошибок человека-оператора, влияния помех в системах связи, сбоев электронного оборудования и по- грешностей систем сканирования и распознавания. Причем в системах, предназначенных для обработки большого объема текстовой информации, например в системах электронного документооборота (СЭД), искажения проявля- ются в основном в виде орфографических ошибок различной кратности (однократные, двукратные, *n*-кратные) [1, 2]. В научных исследованиях, посвященных компьютерной обработке текстовой информации, многократно под- черкивается (главным образом, в виде постановки задач, а не решения проблемы) эффективность использования *n*-граммной модели естественного языка (ЕЯ) для решения задач контроля достоверности передачи и обработки текстов [3]. Однако решение проблемы контроля и коррекции ошибок в текстах на основе *n*-граммной модели, хотя и представляется наиболее перспективным, мало изучено с точки зрения обеспечения качества обработки

текстовой информации, особенно представляемой на узбекском языке.

Следует отметить, что проблема контроля и коррекции ошибок в текстах на основе *n-*граммной модели ЕЯ связана с решением комплекса теоретических и практических задач, среди которых наиболее важными являются: исследование вероятностей появления ошибок для получения априорной базы *n-*грамм; разработка методик оцен- ки достоверности информации при равномерных и неравномерных моделях *n-*кратных искажений; парсинговое моделирование структуры слова на основе словоформ, разработка вероятностных моделей кластеризации и поиска объектов контроля; компьютерная реализация моделей и алгоритмов контроля и коррекции *n-*граммных ошибок, оптимизация параметров функционирования компонентов систем контроля орфографии и оценка качества ее функционирования.

В настоящей работе представлены результаты исследований, направленных на решение указанных задач.

1. **Модели условной вероятности *n-*граммных искажений.** Определение вероятностей *n-*граммных ошибок связано с обработкой большого объема статистических данных и трудоемкими вычислениями, так как важной особенностью *n-*грамм является то, что их число растет экспоненциально относительно длины n. Следовательно, необходимо специальное моделирование процессов вычисления статистики и вероятностей *n-*граммных ошибок. В работе [4] исследованы закономерности распределения ошибок передачи текстовой информации, предложены способы моделирования и алгоритмы для выявления искаженных элементов (букв, слов) в тексте, кластеризации,

поиска, структуризации; получены частотные характеристики *n-*грамм при большом объеме информации, которые применялись в процессах апробации систем контроля и коррекции орфографических ошибок. Результаты прове- денных экспериментальных исследований использовались при установлении закономерностей появления искаже- ний в информации, определении условных вероятностей *n-*граммных ошибок для решения задач генерации и син- теза текстов из речи.

Заметим, что используемые экспериментальные данные получены на основе теоретических положений при допущении о равновероятности *n-*граммных ошибок, что позволило получить простые математические выражения для проведения аналитических исследований. В связи с этим представим равномерную модель *n-*граммных ошибок.

* 1. *Равномерная модель n-граммных ошибок.* Общая вероятность ошибок, обусловленных ошибками челове- ка-оператора, сканирования и распознавания, искажениями в каналах связи, сбоями электронных средств переда- чи и обработки информации, обозначим через *Р*. Процесс перехода *i* -го сообщения в  *j* -е, как правило, задается

стохастической матрицей переходных вероятностей

*P* , которая считается основным показателем при оценке

*i i*

 

достоверности информации в любой системе передачи и обработки данных.

Общая вероятность ошибок при передаче *i* -го сообщения равна

*P**i*  1 *P**i**i*   *P i j* ,

 

 *j* (*i*  *j* )

где

*P*  − вероятность правильного приема *i* -го сообщения. Средняя вероятность ошибки находится осредне-

нием условных вероятностей ошибки по всему ансамблю сообщений:

*i i*

*P*   *P**i*

*i*

 *P i j* .

 *j* (*i*  *j* )

 

(1)

Формула (1) является двумерной моделью оценки вероятности *Р*, связанной с оценкой монограммной вероят-

ности *P* и диграммной вероятности



*i*

*P* . В случае учета статистики трехграмм необходимо исследовать

*i i*

 

вероятности переходов

*i* *j*   , а при статистике *n-*грамм требуется вычислить вероятности набора

*i* ,*i*'' ,L,*in*   *j* ' , *j* '' ,L, *j n* .

* 1. *Математическая модель условных вероятностей n-грамм.* Пусть задан некоторый язык

*L*(*VT* )

с конеч-

ным алфавитом *VT*

 {*wi*} , где *wi*

– отдельный символ, *VT* – множество цепочек (строк) конечной длины, состоя-

щих из символов алфавита *VT* , *n-*грамма на алфавите *VT*

представляет собой цепочку длиной *n*.

Как правило, *n-*грамма может совпадать с каким-либо высказыванием, быть его подстрокой или вообще не

входить в

*L*(*VT* ) . Например, если алфавит – это буквы ЕЯ плюс дефис, а высказывания – это слова ЕЯ, то

*n-*грамма – это последовательность из *n* символов (букв и дефисов), принадлежащая одному слову; если высказы- вания – это тексты, то *n-*грамма – это последовательность из *N* слов одного текста; если алфавит – это морфологи- ческие описания слов ЕЯ плюс знаки пунктуации, а высказывания – это соответствующие фразам и грамматиче- ски допустимые морфологические описания входящих в них слов, то *n-*грамма – это последовательность грамма- тически допустимых описаний *n* подряд стоящих слов.

Обозначим через *C*(*w*)  *C*(*w*1*w*2 …*wn*1*wn* ) число вхождений строки

*w*  *w*1*w*2 …*wn*1*wn*

в совокупность всех текстов рассматриваемого языка. Предположим, что алфавит рассматриваемого языка содер- жит буквы (без учета регистра) и знаки пунктуации, тогда как пробел, переход на новую строку и начало текста – специальные разделители, не входящие в алфавит. Высказывание в таком языке – это неделимая последователь- ность символов

*p*(*w*)  *C*(*w*) .

*C*(*w*\*)

*w*\*

Вероятность

*p*(*w*)

появления *n-*граммы

*w*  *w*1 …*wn*

равна отношению

*C*(*w*)

к общему числу экземпляров

всех встреченных в совокупности *n-*грамм. В частности, для монограмм, т. е. отдельных символов, имеем

*p*(*wi* ) 

*C*(*wi* )

*C*(*w j* )

*w j*

,

где *wi*

– символ алфавита *V* ; числитель – количество вхождений *wi*

в совокупность всех слов, а сумма в знаме-

нателе − общее число символов в ней.

*T*

Если вероятности появления символов в любой позиции цепочки независимы и одинаково распределены, то вероятность *n-*граммы

*n*

*p*(*w*1...*wn*)   *p*(*wi* ) .

*i*1

Это, в частности, означает, что любые перестановки символов строки ность.

*w*  *w*1 …*wn*

имеют одну и ту же вероят-

Если достоверного априорного знания о равенстве распределений символов в разных позициях строки не су-

*j j*

ществует, следует ввести условные вероятности. Тогда, обозначив через позиции строки стоит символ *w*\* , получим условную вероятность строки

*j*

*p*(*w*  *w*\* )

вероятность того, что в *j*-й

*p*(*w*\* …*w*\* )  *p*(*w*

1

*n j j*

 *w*\* *w*

 *w*\**i*  *j*) *p*(*w*  *w*\**i*  *j*) . (2)

Формула (2) служит также априорной основой при построении алгоритмов автоматической кластеризации слов системы контроля орфографии. В связи с этим ниже рассматриваются решение задач кластеризации слов и специфические подходы для получения эффективных алгоритмов кластеризации слов и просмотра строки текста.

*j j i i*

1. **Математическая модель кластеризации слов.** Можно предложить одностороннюю (например, просмотр строки текста слева или справа) и вместе с тем двухстороннюю модель кластеризации слов, где строка текста по- очередно прослеживается и слева, и справа. Установлено, что алгоритм кластеризации на основе односторонней модели позволяет значительно быстрее, без существенных потерь обеспечить выделение слова и разбиение слов на классы. Рассмотрим кластеризацию на основе односторонней модели при просмотре строки текста с левой стороны.

Корпус слов до некоторой степени редуцируется отображением каждого из *Nv*

слов в *Nc*

классы, где

*Nc*  *Nv* .

При этом основным условием является представление *n-*граммной статистики для полученного корпуса классов слов. Для отображения слова в классы данная модель представляется в виде

*w*  *C*  *C*(*w*) ,

где слово *w* может принадлежать только одному классу. В данной работе кластеризация в классы проведена для слов узбекского языка. При этом в качестве критерия оптимизации кластеризации использована мера наибольше- го подобия, определенная в тренировочном множестве. Заметим, что ключевыми моментами кластеризации слов в классы являются парсинговое моделирование структуры слова на основе словоформ [5], выработка методов поис- ка и оценка их вероятностей при принятых моделях.

* 1. *Расчет компонентов вероятностей односторонней модели.* Компонент вероятности односторонней моде- ли классов представляется в виде

*P*(*wi* )  *P*(*wi* / *C*(*wi**n*1 ),…,*C*(*wi*1 )) . (3)

По модели (3) текущее слово обрабатывается в зависимости от предыдущих слов, отображенных в классы. Следо- вательно, вероятность очередного символа строки также задается в зависимости от предшествующих ему (*n*  1)

символов:

*p*(*wn w*1 …*wn*1 ) . Тогда

*p*(*w*1 …*wn*1*wn* )  *p*(*wn w*1 …*wn*1 ) *p*(*w*1 …*wn*1 ) .

В терминах вероятности "быть справа" для триграмм имеем

*p*(*w*1...*wn*)  *p*(*wn* | *w*1...*wn*  1) *p*(*wn*  1 | *w*1...*wn*  2) *p*(*wn*  2 | *w*1...*wn*  3) *p*(*w*2) ,

в общем случае можно записать

 *n* 

*p*(*w*1...*wn*)   *p*(*wk* | *w*1...*wk*  1)  *p*(*w*1) . (4)

 *k* 2 

Введя фиктивный символ "начало" и приняв, что *p*(*w*1 *w*0 ) есть *p*(*w*1) , выражение (4) представим в виде

*n*

*p*(*w*1...*wn*)   *p*(*wk* | *w*1...*wk*  1)

*k* 1

(5)

Таким образом, марковская цепь (*n* 1) -го порядка оказывается моделью *n-*граммы, а задача оценивания ста- тистических параметров *n-*граммы – хорошо изученной задачей оценивания параметров марковской цепи.

Следует отметить, что вследствие наличия множества возможных типичных строк символов значения вероят-

ностей, вычисленные по формуле (5), очень малы и их использование связано с большими трудностями вычисли- тельного характера. Поэтому для упрощения вычислений выражение (4) целесообразно записать в виде

*n*

log *P*(*w*1*w*2 *w*3...*wn* )  log *P*(*w*1 *w*1*w*2 *w*3...*wn* ),

*k* 1

однако для определения (6) необходимы многократные вычисления:

(6)

*n*

*P*   *pi* .

*i*1

Задавая log(*a*  *b*)  log *a*  log(1 *b* / *a*) , вычисляем log *P* по следующему рекурсивному алгоритму: Начало: log *P*  log *p*1

Рекурсия: *a*  max(log *pn* , log *pn*1 )

*b*  min(log *pn* , log *pn*1 )

log *pn*1  *a*  log(1 exp(*b*  *a*))

Конец: log *P*  log *pn* .

Для проведения аналитических исследований эффективности систем контроля орфографии также представляет интерес получение упрощенных оценок вероятностей *n-*грамм.

1. **Упрощенные оценки условных вероятностей *n-*грамм.** Как правило, оценкой вероятности *n-*граммы слу- жит частота ее встречаемости:

*p*ˆ(*w* | *w*

...*w*

)  *f* (*w* | *w*

...*w*

)  *C*(*wi* *n*...*wi* 1*wi* ) .

*i i*  *n i*  1

*i* *n i* 1

*i*

*i i*  *n i*  1

*C*(*w* ...*w w* )

Поскольку частота появления ошибок в виде *n-*грамм представляет случайную величину, частотные характери- стики можно интерполировать для получения их осредненных оценок.

Общая оценка условных вероятностей *n-*грамм также оценивается с учетом частоты их встречаемости:

*p*ˆ(*w* | *w* ...*w*

)  *f* (*w* | *w* ...*w*

)  *C*(*wi**n* …*wi*1*wi* ) ,

*i i**n i*1

*i i**n i*1

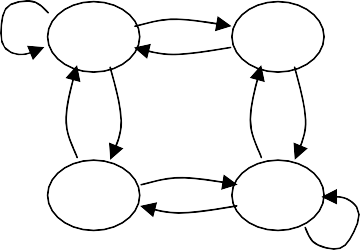
*C*(*wi* )

где *C*(*wi* ) – общее число *n-*грамм, встреченных в последовательности.

В качестве методики получения упрощенной оценки вероятностных переходов предложим упрощенную зна- ково-основанную диграммную модель.

*а б* 4.1. *Диграммная модель ошибок.* Рассмот-

*w* рим диграммную модель, которая требует ве-



*S*1

*S*2

*w*

*S*3

*S*4

роятностей формы *P*(*w w* ) . Обозначим час-

*wi*  *Aw* : *P* (*wi**S*1)

*i j*

*w*44

*w*12

*wi*  *Aw* : *P* (*wi**S*2)

*w*21

*w*22

*Fi j*

*Fi j*

тоты символа или слова через *Fi*, а условные частоты *Fi|j* представим как число следования символа *j* за символом *i*. Тогда оценку макси- мальной вероятности запишем в виде

*P*(*w w* )   .

Рис. 1. Цепь Маркова (*а*) и вероятностные состояния цепи (*б*)

*i j*

 *Fi j Fj*

*i*

Рассмотрим цепь Маркова (рис. 1, *а*), в которой переходы происходят по стрелкам с вероятностями *рij*. На рис. 1, *б* показаны текущие состояния *Si*, выдаваемые символами *i*; причем каждое состояние имеет собственное распределение вероятности.

В данном случае вероятности переходов устанавливаются по формуле

*Fi j* 1

*P*(*wi wj* ) 

.

*Aw*   *Fi j i*

Следует отметить, что вероятности перехода зависят от состояния цепи Маркова, которое является постоян- ным числом. Например, если в момент времени *t* = 0 мы в состоянии *s* с вероятностью перехода *pss*, то вероятность постоянства этого состояния оценивается экспоненциальным разложением

*P*(сост  *s*)  exp(*t* /  )

с характерным временем   1/ log *pss* . Это время прямопропорционально масштабу длины, если модель выдает символы равной длины.

Вероятность переходов между состояниями определим по следующей формуле:

1 *pss*  1 exp(1 /  )  1 /  (  1) .

Большие значения τ исключают переходы в масштабе длины знака и являются желательным поведением системы. Однако если характерное время τ установлено меньшим или равным 1010 знаков, то это не будет подавлять переход.

В случае если известно большее количество данных об индивидуальных частотах символа, то по моделям мо- нограммы лучше определяются вероятности диграмм. Поэтому введем процедуру интерполирования диграммных распределений более простой моделью монограммы:

*P*(*wi wj*

  *Fi*  (1 ) *Fi j* ) ,

*N Fi*

где *N –* общее число символов;  определяется эмпирически.

Модель монограммы с однородным распределением может сглаживать и более сложные модели, например триграммную модель.

4.2. *Триграммная модель ошибок*. С целью упрощения оценки условных вероятностей триграмм будем исполь- зовать линейную интерполяцию

*p*ˆ(*wi* | *wi*2 *wi*1 )  *q*2 *f* (*wi* | *wi*2 *wi*1 )  *q*1 *f* (*wi* | *wi*1 )  *q*0 *f* (*wi* ) ,

где

*f* (*wi* …)

– выборочные оценки, которые определяются следующим образом:

*f* (*w* | *w w*

)  *C*(*wi*2 *wi*1*wi* ) ,

*f* (*w* | *w*

)  *C*(*wi*1*wi* ) ,

*f* (*w* )  *C*(*wi* ) .

*i i*2

*i*1

*C*(*w w* )

*i i*1

*C*(*w* ) *i C*

*i*2 *i*1

*i*1

Здесь *C* – общее число экземпляров всех символов, остальные величины в знаменате- лях – число для соответствующих (*n–*1)-грамм, за которыми следует допустимый в рассмат- риваемом языке символ. В каждом слове это число для (*n–*1)-грамм на единицу меньше, чем для *n-*грамм, в случае если число (*n–*1)- грамм больше нуля, в противном случае это число равно 0.

Для упрощенной вероятностной оценки авторами данной работы предложен метод рекурсивной линейной интерполяции относи- тельных оценок частоты различных порядков

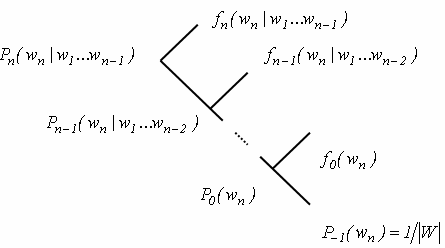


Рис. 2. Рекурсивная линейная интерполяция

*f k* (), *k*  0... *n* . На рис. 2 приведена рекурсивная схема смешивания,

на основе которой запишем выражение для вычисления условных вероятностей

*Pn* (*wn* | *w*1 ,..., *wn*1 )  (*w*1 ,..., *wn* )*Pn*1 (*wn* | *w*1 ,..., *wn*2 )  (1 (*w*1,..., *wn* )) *f n* (*wn* | *w*1 ,..., *wn*1 ) ,

*P*1 (*w*)  uniform(*W* ) ,

где

*w*1 ,…, *wn*1

* контекст порядка *n*, когда предсказано

*wn* ;

*f n* (*wn w*1,…, *wk* )

* относительная частотная оценка

порядка *k* для условной вероятности *Pn* (*wn w*1 ,…, *wk* ) :

*f k* (*wn* | *w*1,…, *wk* )  *C*(*wn* , *w*1,…, *wk* ) *C*(*w*1,…, *wk* ),

*k*  0…*n*,

*C*(*wn* , *w*1,…, *wk* )  

*wk* 1*Wk* 1

… 

*wn* *Wn*

*C*(*wn* , *w*1,…, *wk* , *wk* 1 …*wn* 1 ),

(*w*1,…, *wk* ) [0,1],

*k*  0…*n*

*C*(*w*1,…, *wk* )   *C*(*wn* , *w*1,…, *wk* ),

*w**W*

* коэффициенты интерполяции.

Заметим, что коэффициенты (*w*1 ,…, *wk* )

сгруппированы в эквивалентные классы на основе диапазона, в ко-

торый попадает индекс *C*(*w*1,…, *wk* ) ; для каждого эквивалентного класса диапазоны индекса установлены таким образом, что статистически достаточное число событий (*wn* | *w*1,…, *wk* ) попадает в пределы этого диапазона.

Предложенная выше методика оценки условных вероятностей ошибок в текстах на основе *n-*граммной модели позволяет оценить их значения в виде осредненных характеристик появления однократных, двукратных и трех- кратных ошибок, которые являются важными факторами при оценке качества применения способов контроля дос- товерности текстовой информации.

1. **Оценка достоверности информации.** Поскольку в системах контроля орфографии основным элементом проверки и коррекции является слово текста, при построении таких систем на первый план выдвигаются задачи распознавания слова и его элементов. В [2, 6] разработаны интерполяционные и экстраполяционные алгоритмы распознавания элементов текста, в том числе слова. Ниже рассмотрены методики получения вероятностных моде- лей выделения слов в строке текста в предположении, что распознавание слова осуществляется по указанным ал- горитмам статистического распознавания.
   1. *Вероятностная модель распознавания элементов текста.* Начнем с выделения строки слов

*W*ˆ  arg max *P*( *A* | *W* )*P*(*W* ) ,

&

*W*

где *A* обозначает наблюдаемое слово; *P*( *A* / *W* ) – условная вероятность того, что слово в строке *W* представляется

в виде образа *A* ; *P*(*W* ) – априорная вероятность появления слова в тексте *W*. Исследование заключается в оценке

значения вероятности *P*(*W* ) .

Пусть строка задается набором слов *W*  *w*1, *w*2 ,..., *wn* , тогда по теореме Байеса имеем

*n*

*P*(*W* )   *P*(*wi* | *w*1, *w*2 ,…, *wi*1 ) .

*i*1

Заметим, что пространство параметра

*P*(*wk* | *w*1 , *w*2 ,…, *wk* 1 )

очень широко, причем слова *wi* принадлежат сло-

варю *V* большого размера. Для распознавания представляется предыстория лентного класса, определяемого функцией Ф(*Wk* 1 ) , а также

*Wk*  *w*1 , *w*2 ,…, *wk* 1

в виде эквива-

*n*

*P*(*W* )   *P*(*wk* | Ф(*Wk* 1 )) .

*k* 1

Тогда задача определения вероятности выделения слов сводится к нахождению эквивалентных классификаторов Ф и методов оценки *P*(*wk* | Ф(*Wk* 1 )) .

Поскольку для распознавания слова в тексте предлагается использование *n*-граммной модели языка, функция эквивалентной классификации представляется в виде

Ф(*Wk* 1 )  *wk* *n*1 , *wk* *n*2 ,…, *wk* 1 .

Следует отметить, что определение формы Ф(*Wk* 1 ) предшествует решению задачи оценки *P*(*wk* | Ф(*Wk* 1 )) , яв- ляющейся критерием качества распознавания и соответственно контроля достоверности элементов текста.

* 1. *Оценка качества распознавания слова.* Качество системы контроля орфографии, как правило, определяет-

ся достоверностью распознавания слова на основе словаря словоформ. Поэтому при решении поставленной зада- чи важным моментом является определение показателя ошибки распознавания слова. Для этого находим наиболее

благоприятное слово, произведенное алгоритмом распознавания *W*ˆ

и истинной последовательностью слов. Затем

подсчитывается число неправильных слов *W*ˆ

в общем числе слов в *W* .

Особенность контроля текстовой достоверности заключается в том, что при построении алгоритма распозна- вания и соответственно системы контроля орфографии используется большой объем словарей словоформ и пре- фиксов слов, при этом алгоритм позволяет выделить несоответствующие слова, обеспечить эквивалентную клас- сификацию префикса слова и использовать априорную информацию при предсказании следующего слова.

Как одну из оценок качества распознавания слова можно использовать энтропию основного источника инфор- мации

где *H w*

*N*

*Hw* (*M* )  exp(1 / *N* ln[*PM* (*wk* | *Wk* 1 )]) ,

*k* 1

* энтропия слова в строке; *N* − число слов в общем объеме словаря тестируемого материала.

1. **Парсинговое моделирование структуры слова на основе словоформ.** Аргументы приведенных моделей эквивалентной классификации и оценки качества распознавания определяются на основе изложенного ниже ново- го механизма применения *n-*граммной структурированной модели естественного языка, который включает проце- дуры парсингового кодирования и поиска последовательности контролируемых слов.
   1. *Парсинговое кодирование.* Пусть *W* – предложение длиною *n* слов, к которому добавим в начало  *s*  и в конец  /*s*  , так что получим *w*0  *s*  и *wn*1  *s*  .

Обозначим через *Wk*  *w*0 …*wk*

число *k*-префиксов слова в предложении, тогда *WkTk*

будет *k*-префиксом слова-

парсинга. Для кодирования последовательности слов построим дерево слова-парсинга. Отметим, что *k* -префикс слова-парсинга содержит только те бинарные поддеревья, диапазоны которых полностью включены в *k* -префиксы

слова, за исключением

*w*0  *s*  . Отдельные слова вместе с их позиционными признаками (POS-признак) могут

быть расценены как корневые деревья.

На рис. 3 показан полный парсинг некоторого слова. Схема определяет бинарный парсинг

( *s*  *SB*)(*w*1,*t*1)…(*wntn* )( /*s* , *SE*) , где последовательность *SB/SE* − отличительный POS-признак для

 *s*  /  /*s*  соответственно с ограничениями, что ( /*s* ,*TOP*)

* единственно дозволенный заголовок;

(*w*1,*t*1 )…(*wntn* )( /*s* , *SE*) формирует элемент, возглавляемый ( /*s* ,*TOP*) .

Парсинги определяются, когда ( /*s* ,*TOP*) – заголовок любого элемента, который доминирует (над  /*s*  ),

но не  *s*  .

На рис. 4 представлена схема взаимодействия модулей системы кодирования для построения алгоритма распо- знавания элементов на основе парсингового дерева. Система кодирования состоит из трех модулей:

1. "Предсказатель слова" предсказывает следующее слово *wk* 1 , данное *k*-префиксом слова-парсинга, затем пе- редает управление на "Tаггер";
2. "Taггер" предсказывает POS-признак

*tk* 1

следующего слова, данного *k*-префиксом слова-парсинга, и по-

следнего предсказанного слова *wk* 1 , затем передает управление модулю "Конструктор";

1. "Конструктор" наращивает существующую двоичную расширенную структуру, повторно генерируя перехо- ды, до тех пор пока управление не перейдет к модулю "Предсказатель" по достижении пустого перехода.

Теперь рассмотрим получение оценки вероятностей обмена информацией между модулями парсинговой модели.

* 1. *Вероятностные оценки парсинговой модели.* Обозначим вероятность распознавания последовательности

слов *W* в парсинговой модели через *P*(*W* ,*T* ) , где *T* − дерево полного парсинга. Вероятностная модель должна

быть способной различить желательные и менее желательные парсинги. Для того чтобы получить правильное на- значение вероятности *P*(*W* ,*T* ) , необходимо определить надлежащие условные вероятности каждому переходу.

Вероятность *P*(*W* ,*T* ) последовательности слов *W* и полного парсинга *T* рассчитывается следующим образом:

*n*1

*P*(*W* ,*T* )  [*P*(*w* | *W T* )*P*(*t* | *W T* , *w* )*P*(*T k* | *W T* , *w* ,*t* )] .

*k* 1

*Nk*

*k k* 1 *k* 1

*k k* 1 *k* 1

*k k* 1

*k* 1 *k* 1 *k k*

Здесь

*P*(*T k*

| *W T*

, *w* ,*t* )   *P*( *pk* | *W T* , *w* ,*t* , *pk* … *pk* ) ; *W T*

− (*k* 1) -й префикс слова-парсинга;

*k* 1

*k* 1 *k* 1

*k k i k* 1 *k* 1 *k k* 1

*i*1

*i*1

*k* 1 *k* 1

*w* – слово, предсказанное "Словопредсказателем"; *t* – признак, назначенный для *w* "Таггером"; *T k* − пошаговая

*k k k*

*k* 1

парсинговая структура, которая генерирует *T*  *T*

|| *T k*

, когда парсинговая структура построена на вершине *T*

*k k* 1

*k* 1

*k* 1

и вновь предсказанного слова

*wk* ; запись  обозначает конкатенацию;

*Nk* 1 − число операций, выполняемых

"Конструктором" на позиции *k* входной строки перед передачей управления "Словопредсказателю" ( *Nk* -я опера-

*p*

ция на позиции *k* – нулевой переход, причем

*Nk* представляет собой функцию от *T* );

*k* обозначает *i*-е действие

"Конструктора", выполненное в позиции *k* строки слова, и представляется следующим образом:

*i*

( /*s* ,*TOP* ')

( /*s* ,*TOP*)

Предсказывать слова

( *s* , *SB*)

(*w* \_1,*t* \_1) (*w* \_ *n*,*t* \_ *n*)( /*s* , *SE*)

Примыкать\_(налево, направо)



ПРЕДИКТОР

ТАГГЕР

Нуль

Тег слова

КОНСТРУКТОР

Рис. 3. Полный парсинг Рис. 4. Взаимодействия модулей системы парсингового кодирования

*pk* {(*adjoin*  *left*, *NTag*), (*adjoin*  *right*, *NTag*),(*uniray*, *NTag*)} ,1  *i*  *N* , *pk*  *null*, *i*  *N* .

*i k i k*

Заметим, что каждое (*W T* , *w* ,*t* , *pk* … *pk* ) , *i*  1,...*N* , определяет значащий *k*-префикс слова-парсинга

*WkTk*

*k* 1 *k* 1 *k k* 1

в позиции *k* в предложении.

*i*1 *k*

1. **Алгоритм оптимизации компонентов модели распознавания, контроля достоверности и поиска сло- воформ.** Для гарантирования надлежащей вероятностной модели по набору полных парсингов для любого пред- ложения *W*, вероятностям "Конструктора" и "Предсказателя слова" необходимо задать определенные значения. Набор ограничений на значения вероятностей компонентов различных моделей совместим со следующим алго- ритмом:
2. *P*(*null* | *WkTk* )  1, *ifh* \_{1}.*word*  *s*  и *h* \_{0}  ( /*s* ,*TOP*) , т. е. перед предсказанием  /*s*  гарантиру-

ется, что (  *s*  , *SB*) примыкает к последнему (прошлому) шагу процесса парсинга;

1. *P*((*adjoin*  *right*,*TOP*) | *WkTk* )  1, если *h* \_ 0  ( /*s* ,*TOP*) и *h* \_{1}.*word*  *s*  ;
2. *P*((*adjoin*  *right*,*TOP*) | *WkTk* )  1, если *h* \_ 0  ( /*s* ,*TOP*) и *h* \_{1}.*word*  *s*  .

Шаги 2, 3 гарантируют, что парсинг, произведенный моделью, совместим с определением полного парсинга;

1.  0*s*.*t*.*Wk* 1*Tk* 1 , *P*(*wk*  /*s* | *Wk* 1*Tk* 1 )  . На этом шаге обеспечивается остановка модели. Как только ко- нец символа предложения  /*s*  сгенерирован, модель заканчивает парсинг с вероятностью, равной единице.
   1. *Оптимизация работы "предсказателя"*. Рассмотрим иерархическую схему и алгоритм построения стеков для нахождения нового слова – объекта контроля. Предположим, что каждый стек содержит частичные парсинги – гипотезы, которые были построены одним и тем же числом операций "Предсказателя" и "Конструктора". Частич-

ный парсинг в каждом стеке оценивается согласно принятому критерию ln(*P*(*W* ,*T* ))

вершины.

начиная с самой высокой

На рис. 5 показана схема действий алгоритма, связанных с просмотром нового слова *Wk* 1 . (Здесь *Pk* − макси-

мальное число операций примыкания для *k*-кратного префикса слова; так как дерево двоично, *Pk*  *k* 1 .)

Процедура поиска строится на основе двух параметров:

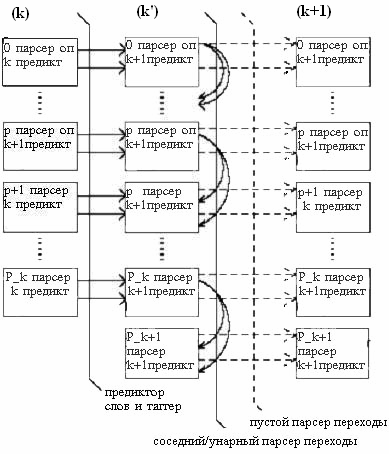


Рис. 5. Цикл расширения поиска

− максимальная глубина стека − максимальное число гипотез, кото- рые стек может содержать в любое данное время;

− порог лог-вероятности − разли- чие между оценками лог-вероятнос- ти наиболее вероятной и наименее вероятной гипотез в любом данном состоянии стека, причем порог лог- вероятности не может быть больше заданного значения.

**Заключение.** Таким образом, теоретические и практические ис- следования проблемы построения компьютерной системы текстовой информации, проведенные с целью разработки методов и алгоритмов контроля и коррекции орфографии на основе *n-*граммной модели есте- ственного языка позволили опреде- лить закономерности распределения *n-*граммных ошибок; оценить досто-

верность информации при равномерных и неравномерных гипотезах *n-*кратных искажений; провести парсинговое кодирование и моделирование структуры слова на основе словоформ; оценить качество распознавания, кластери- зации, поиска элемента текста; моделировать процессы реализации алгоритмов эквивалентной классификации. Полученные вероятностные модели парсингового представления слов, кодирования и поиска позволяют оценить качество распознавания, эффективно моделировать процессы реализации алгоритмов эквивалентной классифика- ции в системах контроля и коррекции орфографических ошибок.

Предложены методы и алгоритмы оптимизации параметров функционирования компонентов системы контро- ля орфографии, которые реализованы в виде самостоятельных программных модулей, соответствующих требова- ниям разработки пакетов прикладных программ. Полученные теоретические положения исследований позволили построить программную систему контроля и коррекции орфографии узбекского языка на основе *n-*граммной мо- дели, которая показала высокое качество функционирования в системах электронного документооборота пред- приятий различных форм собственности.

Список литературы

1. АХАТОВ А. Р. Повышение достоверности информации систем электронного документооборота на прикладных уровнях теле- коммуникационных сетей // Техника и технология. 2008. № 4. С. 25−32.
2. АХАТОВ А. Р., ЖУМАНОВ И. И., ДЖУРАЕВ М. К. Метод проверки орфографических ошибок в текстах на естественных язы- ках // Материалы XIV Междунар. Центрально-Азиатской науч. конф. "Математические методы в технике и технологиях – ММТТ-16", Ташкент (Узбекистан), 22–24 окт. 2003 г. Ташкент: Изд-во Ташкент. химико-технолог. ин-та, 2003. С. 86−89.
3. АХАТОВ А. Р. Алгоритмы программной системы контроля текстовой информации на основе *n*-граммной языковой модели //

Актуальные проблемы современной науки. 2009. № 3. С. 156−161.

1. AKHATOV A. R., JUMANOV I. I., KURBANOV M. M., KARSHIEV Z. A. Use of *N-*gram statistics for checking of the texts transfer quality in intellectual information systems // Рroc. of the 5th World conf. on intelligent systems for industrial automation. Tashkent (Uz- bekistan), 25−27 November, 2008. b-Quadrat Verlag-86916 Kaufering, 2008. P. 153−160.
2. CHARNIAK E. Statistical parsing with a context-free grammar and word statistics // Proc. of the 14th National conf. on artificial intelli- gence, Menlo Park (CA), 19–24 Jul. 1997. AAAI Press/MIT Press, 1997. P. 598–603.
3. АХАТОВ А. Р. Программные методы контроля достоверности информации в структуре пактов передачи данных систем элек- тронного документооборота // Вестн. Сиб. гос. ун-та телекоммуникаций и информатики. 2008. № 2. С. 3−20.

*Камилов Мирзоян Мирзаахмедович – д-р техн. наук, проф. акад. АН Республики Узбекистан, зав. лаб. Ин-та математики и информационных технологий АН РУз;*

*Ахатов Акмал Рустамович – канд. техн. наук, доц. Самаркандского гос. ун-та;*

*тел. (8366) 220-6881, e-mail:* [*akmalar@rambler.ru*](mailto:akmalar@rambler.ru)

Дата поступления – 29.10.2009 г.