УДК 004.512

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

М. М. Камилов, А. Р. Ахатов

Самаркандский государственный университет им. А. Навои, 140104, Самарканд, Узбекистан

Предложен новый подход к построению компьютерной системы передачи и обработки текстовой информации 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 представления и моделирования элементов текста.

The author offers a new approach for construction computer system of transfer and processing the text information on a basis of *n*-gram language model. In article it is stated results of receiving the techniques of definition conditional probabilities of *n*-multiple mistakes in the information. It is developed the ways and algorithms of optimization 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 информационных систем в существенной степени зависит от достоверности передачи сообщений, которая снижается вследствие ошибок человека-оператора, влияния помех в системах связи, сбоев электронного оборудования и погрешностей систем сканирования и распознавания. Причем в системах, предназначенных для обработки большого объема текстовой информации, например в системах электронного документооборота (СЭД), искажения проявляются в основном в виде орфографических ошибок различной кратности (однократные, двукратные, *n*-кратные) [1, 2].

В научных исследованиях, посвященных компьютерной обработке текстовой информации, многократно подчеркивается (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 оценки 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.

**2. 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 проведенных экспериментальных исследований использовались при установлении закономерностей появления искажений в информации, определении условных вероятностей *n*-граммных ошибок для решения задач генерации и синтеза 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.

2.1. *Равномерная модель n-граммных ошибок.* Общая вероятность ошибок, обусловленных ошибками человека-оператора, сканирования и распознавания, искажениями в каналах связи, сбоями электронных средств передачи и обработки информации, обозначим через *P*. Процесс перехода $\alpha_i$-го сообщения в $\alpha_j$-е, как правило, задается стохастической матрицей переходных вероятностей $\left\| P_{\alpha_i \alpha_i} \right\|$, которая считается основным показателем при оценке достоверности информации в любой системе передачи и обработки данных.

Общая вероятность ошибок при передаче $\alpha_i$-го сообщения равна

$$P_{\alpha_i} = 1 - P_{\alpha_i \alpha_i} = \sum_{\alpha_j (\alpha_i \neq \alpha_j)} P_{\alpha_i \alpha_j},$$

где $P_{\alpha_i \alpha_i}$ − вероятность правильного приема $\alpha_i$-го сообщения. Средняя вероятность ошибки находится осреднением условных вероятностей ошибки по всему ансамблю сообщений:

$$P = \sum_{\alpha_i} P_{\alpha_i} \sum_{\alpha_j (\alpha_i \neq \alpha_j)} P_{\alpha_i \alpha_j}. \qquad (1)$$

Формула (1) является двумерной моделью оценки вероятности *P*, связанной с оценкой монограммной вероятности $P_{\alpha_i}$ и диграммной вероятности $P_{\alpha_i \alpha_i}$. В случае учета статистики трехграмм необходимо исследовать вероятности переходов $\alpha_i \alpha_j \to \gamma$, а при статистике *n*-грамм требуется вычислить вероятности набора $\alpha_{i'}, \alpha_{i''}, \cdots, \alpha_{i^n} \to \alpha_{j'}, \alpha_{j''}, \cdots, \alpha_{j^n}$.

2.2. *Математическая модель условных вероятностей n-грамм.* Пусть задан некоторый язык $L(V_T)$ с конечным алфавитом $V_T = \{w^i\}$, где $w^i$ – отдельный символ, $V_T$ – множество цепочек (строк) конечной длины, состоящих из символов алфавита $V_T$, *n*-грамма на алфавите $V_T$ представляет собой цепочку длиной *n*.

Как правило, *n*-грамма может совпадать с каким-либо высказыванием, быть его подстрокой или вообще не входить в $L(V_T)$. 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 высказывания – это тексты, то *n*-грамма – это последовательность из *N* слов одного текста; если алфавит – это морфологические описания слов ЕЯ плюс знаки пунктуации, а высказывания – это соответствующие фразам и грамматически допустимые морфологические описания входящих в них слов, то *n*-грамма – это последовательность грамматически допустимых описаний *n* подряд стоящих слов.

Обозначим через $C(w) = C(w_1 w_2 \ldots w_{n-1} w_n)$ число вхождений строки

$$w = w_1 w_2 \ldots w_{n-1} w_n$$

в совокупность всех текстов рассматриваемого языка. Предположим, что алфавит рассматриваемого языка содержит 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 последовательность символов

$$p(w) = \frac{C(w)}{\sum_{w^*} C(w^*)}.$$

Вероятность $p(w)$ появления $n$-граммы $w = w_1 \dots w_n$ равна отношению $C(w)$ к общему числу экземпляров всех встреченных в совокупности $n$-грамм. В частности, для монограмм, т. е. отдельных символов, имеем

$$p(w^i) = \frac{C(w^i)}{\sum_{w^j} C(w^j)},$$

где $w^i$ – символ алфавита $V_T$; числитель – количество вхождений $w^i$ в совокупность всех слов, а сумма в знаменателе − общее число символов в ней.

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

$$p(w_1 \dots w_n) = \prod_{i=1}^{n} p(w_i).$$

Это, в частности, означает, что любые перестановки символов строки $w = w_1 \dots w_n$ имеют одну и ту же вероятность.

Если достоверного априорного знания о равенстве распределений символов в разных позициях строки не существует, следует ввести условные вероятности. Тогда, обозначив через $p(w_j = w_j^*)$ вероятность того, что в $j$-й позиции строки стоит символ $w_j^*$, получим условную вероятность строки

$$p(w_1^* \dots w_n^*) = p(w_j = w_j^* \,|\, w_j = w_j^* \forall i \neq j)\; p(w_i = w_i^* \forall i \neq j). \qquad (2)$$

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.

**3. 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 поочередно 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 $N_v$ слов в $N_c$ классы, где $N_c < N_v$. При этом основным условием является представление $n$-граммной статистики для полученного корпуса классов слов. Для отображения слова в классы данная модель представляется в виде

$$w \to C = C(w),$$

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 наибольшего подобия, определенная в тренировочном множестве. Заметим, что ключевыми моментами кластеризации слов в классы являются парсинговое моделирование структуры слова на основе словоформ [5], выработка методов поиска и оценка их вероятностей при принятых моделях.

3.1. *Расчет компонентов вероятностей односторонней модели.* Компонент вероятности односторонней модели классов представляется в виде

$$P(w_i) = P(w_i / C(w_{i-n+1}), \dots, C(w_{i-1})). \qquad (3)$$

По модели (3) текущее слово обрабатывается в зависимости от предыдущих слов, отображенных в классы. Следовательно, вероятность очередного символа строки также задается в зависимости от предшествующих ему $(n-1)$ символов: $p(w_n \,|\, w_1 \dots w_{n-1})$. Тогда

$$p(w_1 \dots w_{n-1} w_n) = p(w_n | w_1 \dots w_{n-1}) p(w_1 \dots w_{n-1}) .$$

В терминах вероятности "быть справа" для триграмм имеем

$$p(w_1 \dots w_n) = p(w_n \mid w_1 \dots w_{n-1}) p(w_{n-1} \mid w_1 \dots w_{n-2}) p(w_{n-2} \mid w_1 \dots w_{n-3}) p(w_2) ,$$

в общем случае можно записать

$$p(w_1 \dots w_n) = \left( \prod_{k=2}^{n} p(w_k \mid w_1 \dots w_{k-1}) \right) p(w_1) . \quad (4)$$

Введя фиктивный символ "начало" и приняв, что $p(w_1 | w_0)$ есть $p(w_1)$, выражение (4) представим в виде

$$p(w_1 \dots w_n) = \prod_{k=1}^{n} p(w_k \mid w_1 \dots w_{k-1}) \quad (5)$$

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

Следует отметить, что вследствие наличия множества возможных типичных строк символов значения вероятностей, вычисленные по формуле (5), очень малы и их использование связано с большими трудностями вычислительного характера. Поэтому для упрощения вычислений выражение (4) целесообразно записать в виде

$$\log P(w_1 w_2 w_3 \dots w_n) = \sum_{k=1}^{n} \log P(w_1 | w_1 w_2 w_3 \dots w_n), \quad (6)$$

однако для определения (6) необходимы многократные вычисления:

$$P = \sum_{i=1}^{n} p_i .$$

Задавая $\log(a+b) = \log a + \log(1 + b/a)$, вычисляем $\log P$ по следующему рекурсивному алгоритму:

Начало: $\log P = \log p_1$

Рекурсия: $a = \max(\log p_n, \log p_{n+1})$

$$b = \min(\log p_n, \log p_{n+1})$$

$$\log p_{n+1} = a + \log(1 + \exp(b - a))$$

Конец: $\log P = \log p_n$ .

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

**4. Упрощенные оценки условных вероятностей *n*-грамм.** Как правило, оценкой вероятности *n*-граммы служит частота ее встречаемости:

$$\hat{p}(w_i \mid w_{i-n} \dots w_{i-1}) = f(w_i \mid w_{i-n} \dots w_{i-1}) = \frac{C(w_{i-n} \dots w_{i-1} w_i)}{C(w_{i-n} \dots w_{i-1} w_i)} .$$

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

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

$$\hat{p}(w_i \mid w_{i-n} \dots w_{i-1}) = f(w_i \mid w_{i-n} \dots w_{i-1}) = \frac{C(w_{i-n} \dots w_{i-1} w_i)}{C(w_i)} ,$$

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

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

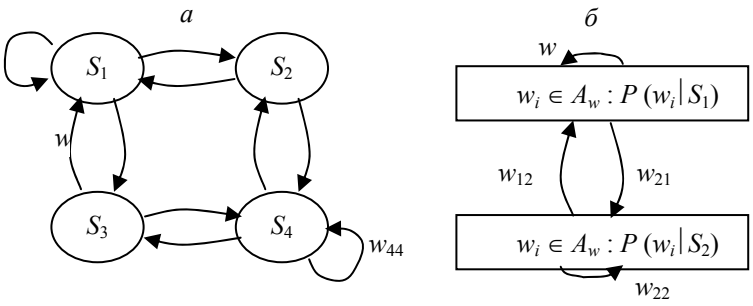


Рис. 1. Цепь Маркова (*а*) и вероятностные состояния цепи (*б*)

4.1. *Диграммная модель ошибок.* Рассмотрим диграммную модель, которая требует вероятностей формы $P(w_i|w_j)$. Обозначим частоты символа или слова через $F_i$, а условные частоты $F_{i|j}$ представим как число следования символа $j$ за символом $i$. Тогда оценку максимальной вероятности запишем в виде

$$P(w_i|w_j) = \frac{F_{i|j}}{\sum_i F_{i|j}} = \frac{F_{i|j}}{F_j} \ .$$

Рассмотрим цепь Маркова (рис. 1, *а*), в которой переходы происходят по стрелкам с вероятностями $p_{ij}$. На рис. 1, *б* показаны текущие состояния $S_i$, выдаваемые символами $\alpha_i$; причем каждое состояние имеет собственное распределение вероятности.

В данном случае вероятности переходов устанавливаются по формуле

$$P(w_i|w_j) = \frac{F_{i|j} + 1}{|A_w| + \sum_i F_{i|j}} \ .$$

Следует отметить, что вероятности перехода зависят от состояния цепи Маркова, которое является постоянным числом. Например, если в момент времени $t = 0$ мы в состоянии $s$ с вероятностью перехода $p_{ss}$, то вероятность постоянства этого состояния оценивается экспоненциальным разложением

$$P(\text{сост} = s) = \exp(-t/\tau)$$

с характерным временем $\tau = -1/\log p_{ss}$. Это время прямопропорционально масштабу длины, если модель выдает символы равной длины.

Вероятность переходов между состояниями определим по следующей формуле:

$$1 - p_{ss} = 1 - \exp(-1/\tau) \approx 1/\tau (\tau << 1) \ .$$

Большие значения τ исключают переходы в масштабе длины знака и являются желательным поведением системы. Однако если характерное время τ установлено меньшим или равным $10^{10}$ знаков, то это не будет подавлять переход.

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

$$P(w_i|w_j = \lambda \frac{F_i}{N} + (1-\lambda)\frac{F_{i|j}}{F_i}) \ ,$$

где $N$ – общее число символов; λ 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 использовать линейную интерполяцию

$$\hat{p}(w_i \mid w_{i-2}w_{i-1}) = q_2 f(w_i \mid w_{i-2}w_{i-1}) + q_1 f(w_i \mid w_{i-1}) + q_0 f(w_i) \ ,$$

где $f(w_i|\ldots)$ – выборочные оценки, которые определяются следующим образом:

$$f(w_i \mid w_{i-2}w_{i-1}) = \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})}, \quad f(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}, \quad f(w_i) = \frac{C(w_i)}{C} \ .$$

Здесь $C$ – общее число экземпляров всех символов, остальные величины в знаменателях – число для соответствующих ($n$–1)-грамм, за которыми следует допустимый в рассматриваемом 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.

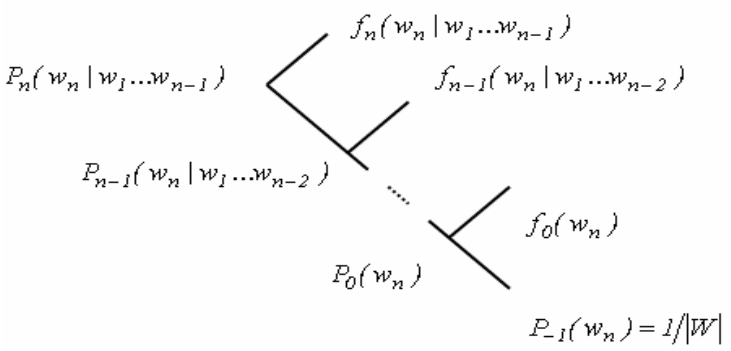


Рис. 2. Рекурсивная линейная интерполяция

Для упрощенной вероятностной оценки авторами данной работы предложен метод рекурсивной линейной интерполяции относительных оценок частоты различных порядков $f_k(\cdot),\ k = 0 \ldots n$ . На рис. 2 приведена рекурсивная схема смешивания, на основе которой запишем выражение для вычисления условных вероятностей

$$P_n(w_n \mid w_1,...,w_{n-1}) = \lambda(w_1,...,w_n)P_{n-1}(w_n \mid w_1,...,w_{n-2}) + (1-\lambda(w_1,...,w_n))f_n(w_n \mid w_1,...,w_{n-1}),$$

$$P_{-1}(w) = \text{uniform}(W),$$

где $w_1,\ldots,w_{n-1}$ – контекст порядка $n$, когда предсказано $w_n$; $f_n(w_n|w_1,\ldots,w_k)$ – относительная частотная оценка порядка $k$ для условной вероятности $P_n(w_n|w_1,\ldots,w_k)$:

$$f_k(w_n \mid w_1,\ldots,w_k) = C(w_n,w_1,\ldots,w_k)/C(w_1,\ldots,w_k),\ \ k = 0\ldots n,$$

$$C(w_n,w_1,\ldots,w_k) = \sum_{w_{k+1}\in W_{k+1}} \ldots \sum_{w_n\in W_n} C(w_n,w_1,\ldots,w_k,w_{k+1}\ldots w_{n-1}),$$

$$C(w_1,\ldots,w_k) = \sum_{w\in W} C(w_n,w_1,\ldots,w_k),$$

$\lambda(w_1,\ldots,w_k) \in [0,1],\ \ k = 0\ldots n$ – коэффициенты интерполяции.

Заметим, что коэффициенты $\lambda(w_1,\ldots,w_k)$ сгруппированы в эквивалентные классы на основе диапазона, в который попадает индекс $C(w_1,\ldots,w_k)$; для каждого эквивалентного класса диапазоны индекса установлены таким образом, что статистически достаточное число событий $(w_n \mid w_1,\ldots,w_k)$ 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 трехкратных ошибок, которые являются важными факторами при оценке качества применения способов контроля достоверности текстовой информации.

**5. 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 моделей выделения слов в строке текста в предположении, что распознавание слова осуществляется по указанным алгоритмам статистического распознавания.

5.1. *Вероятностная модель распознавания элементов текста.* Начнем с выделения строки слов

$$\hat{W} \doteq \arg\max_W P(A \mid W)P(W),$$

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 $W = w_1, w_2, ..., w_n$, тогда по теореме Байеса имеем

$$P(W) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \ldots, w_{i-1}) .$$

Заметим, что пространство параметра $P(w_k \mid w_1, w_2, \ldots, w_{k-1})$ очень широко, причем слова $w_i$ принадлежат словарю $V$ большого размера. Для распознавания представляется предыстория $W_k = w_1, w_2, \ldots, w_{k-1}$ в виде эквивалентного класса, определяемого функцией $\Phi(W_{k-1})$, а также

$$P(W) \cong \prod_{k=1}^{n} P(w_k \mid \Phi(W_{k-1})) .$$

Тогда задача определения вероятности выделения слов сводится к нахождению эквивалентных классификаторов Ф и методов оценки $P(w_k \mid \Phi(W_{k-1}))$.

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

$$\Phi(W_{k-1}) = w_{k-n+1}, w_{k-n+2}, \ldots, w_{k-1} .$$

Следует отметить, что определение формы $\Phi(W_{k-1})$ предшествует решению задачи оценки $P(w_k \mid \Phi(W_{k-1}))$, являющейся критерием качества распознавания и соответственно контроля достоверности элементов текста.

5.2. *Оценка качества распознавания слова.* Качество системы контроля орфографии, как правило, определяется достоверностью распознавания слова на основе словаря словоформ. Поэтому при решении поставленной задачи важным моментом является определение показателя ошибки распознавания слова. Для этого находим наиболее благоприятное слово, произведенное алгоритмом распознавания $\hat{W}$ и истинной последовательностью слов. Затем подсчитывается число неправильных слов $\hat{W}$ в общем числе слов в $W$.

Особенность контроля текстовой достоверности заключается в том, что при построении алгоритма распознавания и соответственно системы контроля орфографии используется большой объем словарей словоформ и префиксов слов, при этом алгоритм позволяет выделить несоответствующие слова, обеспечить эквивалентную классификацию префикса слова и использовать априорную информацию при предсказании следующего слова.

Как одну из оценок качества распознавания слова можно использовать энтропию основного источника информации

$$H_w(M) = \exp(-1/N \sum_{k=1}^{N} \ln[P_M(w_k \mid W_{k-1})]) ,$$

где $H_w$ – энтропия слова в строке; $N$ – число слов в общем объеме словаря тестируемого материала.

**6. Парсинговое моделирование структуры слова на основе словоформ.** Аргументы приведенных моделей эквивалентной классификации и оценки качества распознавания определяются на основе изложенного ниже нового механизма применения *n*-граммной структурированной модели естественного языка, который включает процедуры парсингового кодирования и поиска последовательности контролируемых слов.

6.1. *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 $w_{n+1} =< s >$.

Обозначим через $W_k = w_0 \ldots w_k$ число *k*-префиксов слова в предложении, тогда $W_k T_k$ 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 $(< s > SB)(w_1, t_1) \ldots (w_n t_n)(< /s >, SE)$, где последовательность *SB/SE* – отличительный POS-признак для

$< s > / < /s >$ соответственно с ограничениями, что $(< /s >, TOP)$ – единственно дозволенный заголовок; $(w_1, t_1)\ldots(w_n t_n)(< /s >, SE)$ 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 распознавания элементов на основе парсингового дерева. Система кодирования состоит из трех модулей:

1) "Предсказатель слова" предсказывает следующее слово $w_{k+1}$, данное *k*-префиксом слова-парсинга, затем передает управление на "Таггер";

2) "Таггер" предсказывает POS-признак $t_{k+1}$ следующего слова, данного *k*-префиксом слова-парсинга, и последнего предсказанного слова $w_{k+1}$, затем передает управление модулю "Конструктор";

3) "Конструктор" наращивает существующую двоичную расширенную структуру, повторно генерируя переходы, 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.

6.2. *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 назначение 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:

$$P(W,T) = \prod_{k=1}^{n+1}[P(w_k \mid W_{k-1}T_{k-1})P(t_k \mid W_{k-1}T_{k-1}, w_k)P(T_{k-1}^k \mid W_{k-1}T_{k-1}, w_k, t_k)].$$

Здесь $P(T_{k-1}^k \mid W_{k-1}T_{k-1}, w_k, t_k) = \prod_{i=1}^{N_k} P(p_i^k \mid W_{k-1}T_{k-1}, w_k, t_k, p_1^k \ldots p_{i-1}^k)$; $W_{k-1}T_{k-1}$ – $(k-1)$-й префикс слова-парсинга; $w_k$ – слово, предсказанное "Словопредсказателем"; $t_k$ – признак, назначенный для $w_k$ "Таггером"; $T_{k-1}^k$ – пошаговая парсинговая структура, которая генерирует $T_k = T_{k-1} \| T_{k-1}^k$, когда парсинговая структура построена на вершине $T_{k-1}$ и вновь предсказанного слова $w_k$; запись $\|$ обозначает конкатенацию; $N_{k-1}$ – число операций, выполняемых "Конструктором" на позиции *k* входной строки перед передачей управления "Словопредсказателю" ($N_k$-я операция на позиции *k* – нулевой переход, причем $N_k$ представляет собой функцию от $T$); $p_i^k$ обозначает *i*-е действие "Конструктора", выполненное в позиции *k* строки слова, и представляется следующим образом:

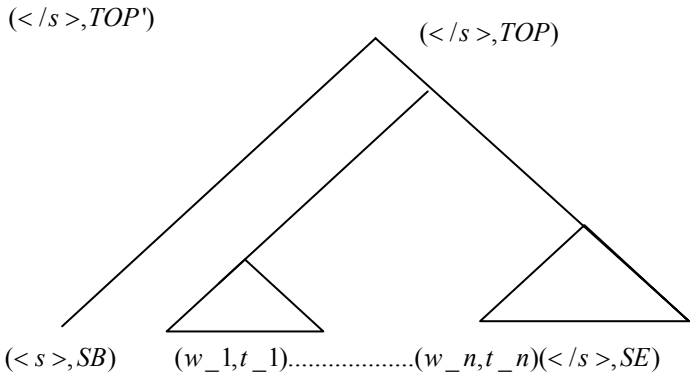


Рис. 3. Полный парсинг

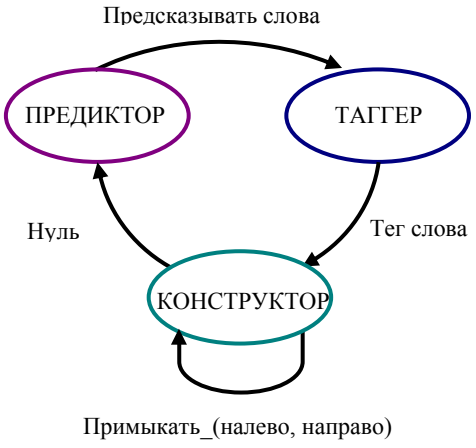


Рис. 4. Взаимодействия модулей системы парсингового кодирования

$$p_i^k \in \{(adjoin-left, NTag),(adjoin-right, NTag),(uniray, NTag)\}, 1 \le i < N_k, \quad p_i^k = null, \quad i = N_k.$$

Заметим, что каждое $(W_{k-1}T_{k-1}, w_k, t_k, p_1^k \dots p_{i-1}^k)$, $i = 1,\dots N_k$, определяет значащий *k*-префикс слова-парсинга $W_kT_k$ в позиции *k* в предложении.

**7. Алгоритм оптимизации компонентов модели распознавания, контроля достоверности и поиска словоформ.** Для гарантирования надлежащей вероятностной модели по набору полных парсингов для любого предложения *W*, вероятностям "Конструктора" и "Предсказателя слова" необходимо задать определенные значения. Набор ограничений на значения вероятностей компонентов различных моделей совместим со следующим алгоритмом:

1. $P(null \mid W_kT_k) = 1$, $if h\_\{-1\}.word =< s >$ и $h\_\{0\} \ne (< /s >, TOP')$, т. е. перед предсказанием $< /s >$ гарантируется, что ($< S >$, *SB*) примыкает к последнему (прошлому) шагу процесса парсинга;

2. $P((adjoin-right, TOP) \mid W_kT_k) = 1$, если $h\_0 = (< /s >, TOP')$ и $h\_\{-1\}.word =< s >$;

3. $P((adjoin-right, TOP') \mid W_kT_k) = 1$, если $h\_0 = (< /s >, TOP')$ и $h\_\{-1\}.word \ne < s >$.

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

4. $\exists \in > 0 s.t. \forall W_{k-1}T_{k-1}, P(w_k =< /s > \mid W_{k-1}T_{k-1}) \ge \in$. На этом шаге обеспечивается остановка модели. Как только конец символа предложения $< /s >$ сгенерирован, модель заканчивает парсинг с вероятностью, равной единице.

7.1. *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". Частичный парсинг в каждом стеке оценивается согласно принятому критерию $\ln(P(W,T))$ начиная с самой высокой вершины.

На рис. 5 показана схема действий алгоритма, связанных с просмотром нового слова $W_{k+1}$. (Здесь $P_k$ − максимальное число операций примыкания для *k*-кратного префикса слова; так как дерево двоично, $P_k = k-1$.)

Процедура поиска строится на основе двух параметров:

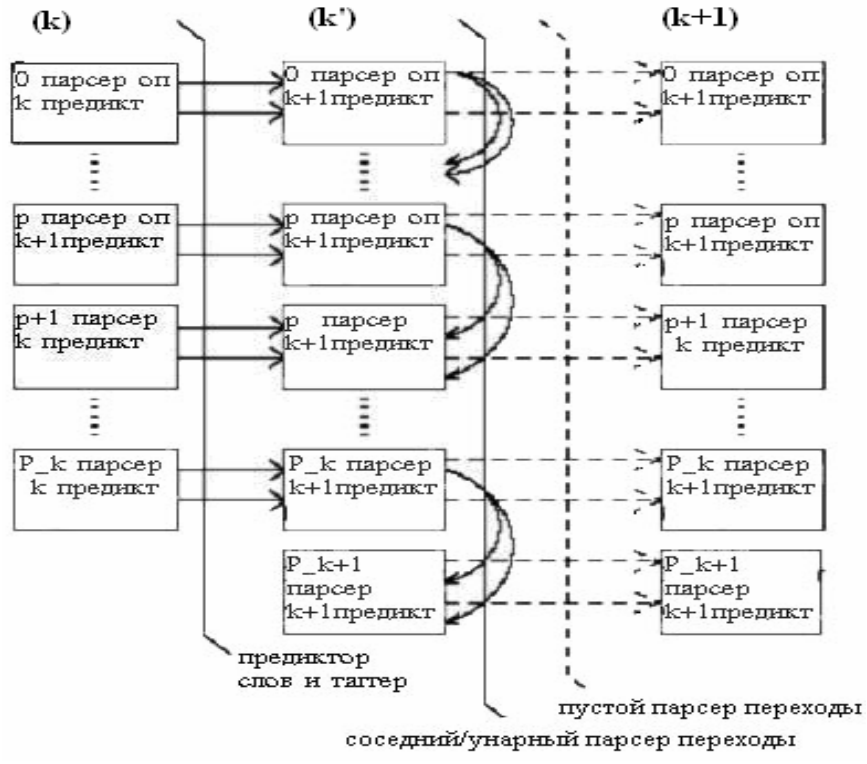


Рис. 5. Цикл расширения поиска

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

− порог лог-вероятности − различие между оценками лог-вероятности 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 исследования проблемы построения компьютерной системы текстовой информации, проведенные с целью разработки методов и алгоритмов контроля и коррекции орфографии на основе *n*-граммной модели естественного языка позволили определить 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, кластеризации, 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 классификации в системах контроля и коррекции орфографических ошибок.

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

## Список литературы

1. Ахатов А. Р. Повышение достоверности информации систем электронного документооборота на прикладных уровнях телекоммуникационных 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 языках // 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.

4. 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 // Proc. of the 5th World conf. on intelligent systems for industrial automation. Tashkent (Uzbekistan), 25−27 November, 2008. b-Quadrat Verlag-86916 Kaufering, 2008. P. 153−160.

5. Charniak E. Statistical parsing with a context-free grammar and word statistics // Proc. of the 14th National conf. on artificial intelligence, Menlo Park (CA), 19–24 Jul. 1997. AAAI Press/MIT Press, 1997. P. 598–603.

6. Ахатов А. Р. Программные методы контроля достоверности информации в структуре пактов передачи данных систем электронного 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*

Дата поступления – 29.10.2009 г.