Информационные процессы, Том 8, № 2, 2008, стр. 138–148.

Модель условного вероятностного автомата в описании ошибок распознавания символов

М.М. Лепешкин

Поступила в редколлегию 23.06.2008 г.

Аннотация—Предлагаются обобщенные модели условного квазивероятностного автомата (УКВ-автомата) и условного вероятностного автомата (УВ-автомата) для описания дискретных каналов с вставками, выпадениями и замещениями символов. Доказывается критерий порождения условного распределения вероятностей и эквивалентное определение УВ-автомата. Используя понятие эквивалентности УКВ-автоматов сужается класс УВ-автоматов, имеющих практическую значимость, до УВ-автоматов без недостижимых состояний. Предлагается способ описания ошибок распознавания символов на основе УВ-автомата, приводятся модели ошибок без памяти, группирования ошибок и Марковская модель ошибок.

1. ВВЕДЕНИЕ

Для моделирования ошибок, возникающих в процессе оптического распознавания символов применяются дискретные каналы с ошибками [1, 2]. При этом ошибки разделяются на три класса: вставки, выпадения и замещения символов [1, 3, 4]. Если представить поступающие входные данные в виде слова некоторого конечного алфавита A, а искаженные ошибками данные в виде слова конечного алфавита B, то процесс внесения ошибок может быть промоделирован дискретным каналом с ошибками в виде вставок, выпадений и замещений символов. Такие преобразования символов в литературе имеют название операций редактирования [3]. Таким образом, множество операций редактирования можно представить в виде: $E = E_s \times E_d \times E_i$, где $E_s = A \times B$– операции замещения, $E_d = A \times \{\lambda\}$ – операции удаления (выпадения), а $E_i = \{\lambda\} \times B$ – операции вставки символов.

Для сравнения слов $x \in A^*$ и $y \in B^*$ в условиях подобных ошибок для случая $|A| = |B| = 2$ Левенштейном была предложена функция расстояния, равная минимальному количеству вставок, выпадений или замещений, переводящих слово x в слово y [5]. Однако данная функция расстояния не учитывает вероятности появления ошибок различных типов, а также зависимость между ошибками. Расстояние Левенштейна можно обобщить путем придания веса каждой операции редактирования. Такое расстояние называют расстоянием редактирования.

В работе [3] в качестве весов операций редактирования предлагается использовать отрицательные логарифмы их вероятностей в предположении, что отдельные операции редактирования независимы. При этом учитывается только один (наиболее вероятный) вариант преобразования слов. Для того чтобы учесть при вычислении функции расстояния всевозможные варианты преобразования слов, в работе [3] предлагается функция стохастического расстояния, равная отрицательному логарифму совместной вероятности появления слов x и y:

$$d^s(x, y) = -\log P(x, y). \tag{1}$$

Для описания дискретного канала при этом используется модель вероятностного автомата [6, 7, 8] без памяти, а для определения параметров (обучения) модели предлагается использовать экспериментальные данные. При этом возникает зависимость оценок параметров от распределения обучающей выборки. Для снижения влияния распределения обучающей выборки в

определении стохастического расстояния может быть использовано условное распределение [9]:

$$d^{cs}(x,y) = -\log P(y|x). \tag{2}$$

Данную функцию расстояния будем называть условным стохастическим расстоянием. Условное распределение $P(y|x)$ в этом случае в работе [9] моделируется автоматом без памяти с помеченными (взвешенными) переходами, где каждому ребру на графе переходов автомата сопоставлена условная вероятность выхода при условии входа.

Все приведенные модели основываются на предположении о независимости отдельных операций редактирования, что в общем случае не верно при моделировании ошибок распознавания, которые могут группироваться, а также зависеть от расположения на изображении. Поэтому требуется построение более общей модели ошибок, позволяющей дать математическое описание зависимостям ошибок.

В пункте 2 настоящей статьи дается определение УВ-автомата и УКВ-автомата. В пункте 3 формулируются критерий порождения условного распределения вероятностей УКВ-автоматом, эквивалентное определение УВ-автомата, при помощи понятия эквивалентности снижается мощность класса различимых автоматов, а также предлагается вычислительная формула для условной вероятности $P(y|x)$. В пункте 4 описываются три модели ошибок распознавания: с независимыми ошибками, группированием ошибок и Марковская модель ошибок. Доказательства основных результатов приводятся в пункте 5.

2. ОПРЕДЕЛЕНИЕ УСЛОВНОГО ВЕРОЯТНОСТНОГО АВТОМАТА

Определение 1. Условным квазивероятностным автоматом (УКВ-автоматом), будем называть следующий автомат с помеченными переходами:

$$\pi = \left\langle Q, \hat{A}, \hat{B}, T, P_I, P_F, P_T \right\rangle, \tag{3}$$

где:

- Q– множество состояний;
- A– входной алфавит, $\hat{A} = A \cup \{\lambda\}$;
- B – выходной алфавит, $\hat{B} = B \cup \{\lambda\}$;
- $T \subseteq Q \times \hat{A} \times \hat{B} \times Q$ – множество переходов;
- $P_I : Q \to [0,1]$ – вероятности того, что состояние является начальным;
- $P_F : Q \to [0,1]$ – вероятности того, что состояние является конечным;
- $P_T : T \to [0,1]$ – вероятности переходов.

При этом $\forall q \in Q, \forall a \in A$ выполняются условия нормировки:

$$\begin{aligned} &\sum_{q\in Q} P_I(q) = 1; \\ &P_F(q) + \sum_{q'\in Q, b\in B} P_T(q,\lambda,b,q') = 1; \\ &\sum_{q'\in Q, b\in B} P_T(q,a,b,q') + \sum_{q'\in Q, b\in B} P_T(q,\lambda,b,q') + \sum_{q'\in Q} P_T(q,a,\lambda,q') = 1. \end{aligned} \tag{4}$$

Символ λ здесь и далее обозначает место вставки или выпадения символов. Функцию $P_T(q,a,b,q')$ можно понимать как вероятность того, что автомат перейдет в состояние q' с появлением на

выходе символа b при условии, что он находится в состоянии q и на вход подается символ a. Однако следует отметить, что функции P_F и P_T не являются распределениями вероятностей.

Если для некоторого состояния $q \quad P_I(q) > 0$, то такое состояние называют начальным. Если же $P_F(q) > 0$, то состояние называют конечным. Будем далее называть состояние $q \in Q$ достижимым из состояния $s \in Q$, если $q = s$ или найдутся два слова некоторой длины $n \quad x^n \in \hat{A}^*$, $y^n \in \hat{B}^*$ и путь $\theta = (q_0, x_1, y_1, q_1, x_2, y_2, q_2, ..., x_n, y_n, q_n) \in \Theta(x, y)$, переводящий x^n в y^n, для которого $q_0 = s$, $q_n = q$ и $\prod_{i=1}^{n} P_T(q_{i-1}, x_i, y_i, q_i) > 0$. При этом состояние s может не быть начальным, а состояние q – конечным. Будем называть состояние $q \in Q$ достижимым, если оно достижимо из некоторого начального состояния.

Определим функции, которые будем называть условной вероятностью преобразования слов и условной вероятностью пути соответственно:

$$\begin{aligned} P_\pi(y|x) &= \sum_{\{\theta \in \Theta(z) | \upsilon(z) = (x,y)\}} P_\pi(\theta); \\ P_\pi(\theta) &= P_I(q_0) \left(\prod_{i=1}^{n} P_T(q_{i-1}, \hat{x}_i, \hat{y}_i, q_i) \right) P_F(q_n). \end{aligned} \tag{5}$$

где $z^n \in E^*$, $\upsilon(z^n) = (x, y) \in A^* \times B^*$ такие, что y может быть получен из x путем последовательного применения операций редактирования из последовательности z^n.

Заметим, что в общем случае для УКВ-автомата функция $P_\pi(.|x)$ может не являться распределением вероятностей. Поэтому ограничим класс рассматриваемых автоматов, только теми, которые порождают условные распределения выходных слов на B^*.

Определение 2. УКВ-автомат (3) с условиями нормировки (4) будем называть условным вероятностным автоматом (УВ-автоматом), если для всех $x \in A^*$ функция $P_\pi(.|x)$ является распределением вероятностей.

Определение 3. Будем говорить, что УКВ-автоматы π_1 и π_2 с одинаковыми входным и выходным алфавитами эквивалентны и обозначать $\pi_1 \sim \pi_2$, если для всех $x \in A^*$ и $y \in B^*$ выполняется: $P_{\pi_1}(y|x) = P_{\pi_2}(y|x)$.

Корректность данного определения следует из выполнения свойств рефлексивности, симметричности и транзитивности для отношения равенства действительных чисел. Очевидно, что если π_1 является УВ-автоматом, то любой эквивалентный ему π_2 также является УВ-автоматом.

Таким образом, эквивалентные УВ-автоматы порождают одинаковые условные распределения вероятностей на множестве выходных слов. Поэтому для изучения свойств данных автоматов достаточно исследовать представителей классов эквивалентности.

3. СВОЙСТВА УКВ- И УВ-АВТОМАТОВ

В работе [10] приводится пример УВ-автомата с двумя состояниями, на котором показывается корректность определения распределения вероятностей по формуле (5) для одного частного случая, рассматриваемого в примере. В общем случае требуется доказательство того, что данное выражение определяет распределение вероятностей.

По УКВ-автомату π и входной строке x^t построим автономный автомат $\mathfrak{A}(\pi, x^t)$, состояния которого хранят состояния исходного автомата π и префиксы входного распределения

следующим образом:

$$
\begin{aligned}
&\mathfrak{A}\left(\pi, x^t\right) = \left\langle Q \times \left(A^{\leqslant t} \cup \{\lambda\}\right), B, T', P'_I, P'_F, P'_T \right\rangle; \\
&T' = \left\{ \left((q, x^r), b, \left(q', x^{r+1}\right)\right) \middle| \left(q, x_{r+1}, b, q'\right) \in T, r \in \overline{1, t-1} \right\} \cup \\
&\quad \cup \left\{ \left((q, x^r), b, \left(q', x^r\right)\right) \middle| \left(q, \lambda, b, q'\right) \in T, r \in \overline{1, t-1} \right\} \cup \\
&\quad \cup \left\{ \left((q, x^r), \lambda, \left(q', x^{r+1}\right)\right) \middle| \left(q, x_{r+1}, \lambda, q'\right) \in T, r \in \overline{1, t-1} \right\}; \\
&\forall q \in Q \; P'_I(q, \lambda) = P_I(q), \; \forall q \in Q, \; \forall r \in \overline{1, t} \; P'_I(q, x^r) = 0; \\
&\forall q \in Q \; P'_F\left(q, x^t\right) = P_F(q), \; \forall q \in Q, \; \forall r \in \overline{1, t-1} \; P'_F(q, x^r) = 0; \\
&P'_T\left((q, x^r), b, \left(q', x^{r+1}\right)\right) = P_T\left(q, x_{r+1}, b, q'\right); \\
&P'_T\left((q, x^r), b, \left(q', x^r\right)\right) = P_T\left(q, \lambda, b, q'\right); \\
&P'_T\left((q, x^r), \lambda, \left(q', x^{r+1}\right)\right) = P_T\left(q, x_{r+1}, \lambda, q'\right).
\end{aligned} \tag{6}
$$

Для удобства в дальнейшем будем считать, что слово нулевой длины эквивалентно при записи символу λ, то есть $x^0 = \lambda$. Корректность определения автомата $\mathfrak{A}\left(\pi, x^t\right)$ доказывается в утверждении 1.

Утверждение 1. *Для автономного автомата с помеченными переходами*

$$\mathfrak{A}\left(\pi, x^t\right) = \left\langle Q', B, T', P'_I, P'_F, P'_T \right\rangle,$$

построенного по процедуре (6) из автомата π, условия нормировки для вероятностного автомата

$$\sum_{s \in Q'} P'_I(s) = 1, \quad \forall s \in Q' \;\; P'_F(s) + \sum_{s' \in Q', b \in B} P'_T\left(s, b, s'\right) = 1 \tag{7}$$

выполняются тогда и только тогда, когда для УКВ-автомата π выполняются условия нормировки (4).

Определение 4. Будем говорить, что в УКВ-автомате π состояние q достижимо из s при помощи $(x_{r+1}, ..., x_u)$, возможно нулевой длины, если вероятность перехода автомата из состояния s в состояние q при подаче на вход некоторой последовательности $(\hat{a}_1, ..., \hat{a}_n)$, такой что $n \geqslant u - r$ и для $1 \leqslant i_{r+1} < i_{r+2} < ... < i_u \leqslant n \quad \hat{a}_{i_j} = x_j$, а $\hat{a}_i = \lambda$ для $i \neq i_j$, $j \in \overline{r+1, u}$, больше нуля, то есть существуют соответствующий путь, имеющий ненулевую вероятность. Если состояние q в автомате π достижимо из некоторого начального состояния, то такое состояние будем называть достижимым.

Заметим, что последовательность $(x_{r+1}, ..., x_u)$ получается из $(\hat{a}_1, ..., \hat{a}_n)$ удалением всех символов λ. Поэтому последовательность $(x_{r+1}, ..., x_u)$ может иметь нулевую длину в случае, когда $u = r$.

Утверждение 2. *Для того, чтобы в автомате $\mathfrak{A}\left(\pi, x^t\right)$ состояние (q, x^u) было достижимо из (s, x^r) необходимо и достаточно, чтобы в автомате π состояние q было достижимо из s при помощи некоторого подслова $(x_{r+1}, ..., x_u)$ слова x^t (возможно нулевой длины).*

Нулевая длина подслова слова x^t означает, что либо соответствующие состояния равны, либо одно состояние достижимо из другого при помощи входного слова, содержащего только символы λ.

При помощи утверждений 1 и 2 можно свести доказательство корректности определения условной вероятности преобразования слов (5) к уже доказанному в работе [11] критерию

корректности для автономного вероятностного автомата. Общая формулировка критерия корректности определения условной вероятности преобразования слов приводится в следующей теореме.

Теорема 1. *(критерий порождения условного распределения вероятностей). УКВ-автомат π (3) порождает условное распределение $P_\pi(.|x)$, задаваемое выражением (5) для $x^t \in A^*$ тогда и только тогда, когда для всякого $r \in \overline{0,t}$ из любого состояния, достижимого при помощи слова $(x_1, ..., x_r)$, возможно, нулевой длины, конечное состояние достижимо при помощи слова $(x_{r+1}, ..., x_t)$, возможно, нулевой длины.*

В качестве естественного следствия из данной теоремы можно сформулировать критерий и эквивалентное определение УВ-автомата.

Следствие. *(эквивалентное определение УВ-автомата). УКВ-автомат π является УВ-автоматом тогда и только тогда, когда для любого $x^t \in A^*$, для всякого $r \in \overline{0,t}$ из любого состояния, достижимого при помощи слова $(x_1, ..., x_r)$, возможно, нулевой длины, конечное состояние достижимо при помощи слова $(x_{r+1}, ..., x_t)$, возможно, нулевой длины.*

На основании следствия из теоремы 3 при проверке корректности УВ-автомата вычисление $P_\pi(y|x)$ для всех $x \in A^*$ и $y \in B^*$ можно заменить обходом графа переходов автомата.

Теорема 3 и следствие из нее накладывают ограничения на достижимые состояния, обеспечивающие порождение автоматом условного распределения вероятностей. Однако в них не рассматриваются недостижимые состояния.

Теорема 2. *Для любого УВ-автомата $\pi = \left\langle Q_\pi, \hat{A}, \hat{B}, T_\pi, P_{\pi,I}, P_{\pi,F}, P_{\pi,T} \right\rangle$ существует эквивалентный ему УВ-автомат $\chi_\pi = \left\langle Q_\chi, \hat{A}, \hat{B}, T_\chi, P_{\chi,I}, P_{\chi,F}, P_{\chi,T} \right\rangle$ не имеющий недостижимых состояний.*

Таким образом, недостижимые состояния в УВ-автомате не влияют на его характеристики. Поэтому при изучении УВ-автоматов достаточно рассматривать только автоматы, не имеющие недостижимых состояний.

Используя приведенный в настоящей работе способ построения вероятностного автомата $\mathfrak{A}(\pi, x^t)$ и основываясь на вычислительной формуле для автономного вероятностного автомата [12] и алгоритме вычисления вероятности наиболее вероятного пути в условном вероятностном автомате [10] значение$P_\pi(y^v|x^t)$ может быть вычислено при помощи следующей рекуррентной

формулы:

$$\begin{aligned}
&\alpha_{x,y}(0,0,q) = P_I(q);\\
&\alpha_{x,y}(0,j,q) = \sum_{s\in Q} \alpha_{x,y}(0,j-1,s)\, P_T(s,\lambda,y_j,q),\ j\in\overline{1,v};\\
&\alpha_{x,y}(i,0,q) = \sum_{s\in Q} \alpha_{x,y}(i-1,0,s)\, P_T(s,x_i,\lambda,q),\ i\in\overline{1,t};\\
&\alpha(i,j,q) = \sum_{s\in Q} \alpha_{x,y}(i-1,j-1,s)\, P_T(s,x_i,y_j,q) +\\
&\qquad + \sum_{s\in Q} \alpha_{x,y}(i-1,j,s)\, P_T(s,x_i,\lambda,q) +\\
&\qquad + \sum_{s\in Q} \alpha_{x,y}(i,j-1,s)\, P_T(s,\lambda,y_j,q),\ i\in\overline{1,t}, j\in\overline{1,v};\\
&P_\pi(y|x) = \sum_{q\in Q} \alpha_{x,y}(t,v,q)\, P_F(q).
\end{aligned} \tag{8}$$

Алгоритм вычисления условной вероятности по рекуррентной формуле (8) имеет сложность $O\left(t\cdot v\cdot n^2\right)$, где $n = |Q|$. Для некоторых частных случаев УВ-автомата с фиксированной структурой переходов вычислительная сложность может быть уменьшена. Например, если допустить единственность начального состояния, а также что входной символ и текущее состояние автомата однозначно определяют следующее состояние, то в формуле (8) не требуется суммирование по всем состояниям. При этом сложность снижается до $O(t\cdot v)$, которая является минимально возможной, согласно [3, 9].

4. МОДЕЛЬ ОШИБОК РАСПОЗНАВАНИЯ НА ОСНОВЕ УВ-АВТОМАТА

Для задания модели ошибок распознавания в форме дискретного канала необходимо определить следующее множество распределений вероятностей:

$$\{P(.|x)\,|x\in A^*\}, \tag{9}$$

где $P(.|x)$ – распределение вероятностей выходных слов (результатов распознавания) на множестве B^* при условии поступления на вход (предъявления к распознаванию) слова $x\in A^*$.

Допустим априорно или на основании экспериментальных данных известны некоторые характеристики данных распределений или характеристики ошибок. Тогда в зависимости от данных характеристик выбирается структура переходов, отражающая зависимости ошибок, вероятности элементарных ошибок и безошибочного распознавания отдельных символов $P_T(q,a,b,q')$, вероятности начальных и конечных состояний $P_I(q)$ и $P_F(q)$. Рассмотрим некоторые характерные случаи.

Независимые ошибки. В условии независимости моделируемых ошибок автоматная память не требуется, поэтому $|Q| = 1$, а все переходы начинаются и заканчиваются в единственном состоянии, то есть можно записать: $T\subseteq \hat{A}\times\hat{B}$. При этом условия нормировки можно переписать в следующем виде:

$$\begin{aligned}
&P_F + \sum_{b\in B} P_T(\lambda,b) = 1;\\
&\sum_{b\in B} P_T(a,b) + \sum_{b\in B} P_T(\lambda,b) + P_T(a,\lambda) = 1.
\end{aligned} \tag{10}$$

Данная модель ошибок без памяти подробно описана в работах [3, 9].

Простое группирование ошибок. Для моделирования группирующихся ошибок распознавания предположим, что вероятность возникновения ошибки в данном символе входного слова зависит только от наличия ошибки на предыдущем шаге автомата и не зависит от характера ошибки. В этом случае множество состояний состоит из двух элементов:

$$Q = \{0, e\},$$

где 0 – отсутствие ошибки на предыдущем шаге, а e – наличие ошибки.

Запишем условия группирования ошибок следующим образом:

$$\begin{aligned}
&\forall a \in A \quad P_T(0,a,a,0) > \sum_{b \in \hat{B}: b \neq a} P_T(0,a,b,e);\\
&\forall a \in A \quad P_T(e,a,a,0) < \sum_{b \in \hat{B}: b \neq a} P_T(e,a,b,e);\\
&\forall a \in A, b \in \hat{B} : b \neq a \quad P_T(0,a,a,e) = P_T(e,a,a,e) = P_T(0,a,b,0) = P_T(e,a,b,0) = 0.
\end{aligned} \tag{11}$$

Данные условия обеспечивают появление ошибки с большей вероятностью, в случае если в предыдущем символе также наблюдалась ошибка, таким образо, моделируется группирование ошибок. При этом, должны выполняться условия нормировки (4), а оба состояния должны быть достижимы и являться конечными, тогда приведенный автомат будет УВ-автоматом.

Марковская модель ошибок. Данная модель основывается на предположении о зависимости ошибок распознавания только от ошибки, произошедшей на предыдущем шаге. То есть, для произвольных $x^t \in A^*$, $y^v \in B^*$ и любой последовательности операций редактирования $z^n \in E^*$ при $n \geqslant t$ и $n \geqslant v$ выполняется:

$$\begin{aligned}
&P(y_j|x_i, z_{k-1}) = P(y_j|x_i, z_{k-1}, ..., z_1), \ z_k = (x_i, y_j);\\
&P(\lambda|x_i, z_{k-1}) = P(\lambda|x_i, z_{k-1}, ..., z_1), \ z_k = (x_i, \lambda);\\
&P(y_j|\lambda, z_{k-1}) = P(y_j|\lambda, z_{k-1}, ..., z_1), \ z_k = (\lambda, y_j).
\end{aligned} \tag{12}$$

Используя терминологию УВ-автоматов, можно переписать выражение (12) с применением функции P_T. Для этого рассмотрим автомат с входным алфавитом A, выходным алфавитом B и множеством состояний

$$Q = E = \left\{(a,b) \,|a \in \hat{A}, b \in \hat{B}\right\} \setminus \{(\lambda,\lambda)\}.$$

Тогда условие (12) представляется в следующем виде:

$$\begin{aligned}
&\forall a, a' \in A^*, b, b' \in B^* \quad P_T\left(\left(a', b'\right), b, a, (a,b)\right) \geqslant 0;\\
&\forall a, a', a'' \in A^*, b, b', b'' \in B^* : a \neq a'' \vee b \neq b'' \quad P_T\left(\left(a', b'\right), b, a, \left(a'', b''\right)\right) = 0.
\end{aligned} \tag{13}$$

Условие (13) обеспечивает ненулевую вероятность перехода автомата только в состояние, описывающее операцию редактирования, выбранную автоматом на данном шаге. Таким образом, на следующем шаге распределение вероятностей выходных символов будет определяться только входом и данной операцией редактирования, сохраненной в состоянии автомата, что обеспечивает выполнение соотношения (12).

Также для автомата должны выполняться условия нормировки (4), а из любого достижимого состояния должно быть достижимо конечное состояние, что гарантирует данному автомату соответствие определению УВ-автомата.

5. ДОКАЗАТЕЛЬСТВА ОСНОВНЫХ РЕЗУЛЬТАТОВ

Доказательство утверждения 1. Пусть выполняются условия нормировки (4) для исходного автомата π. Докажем сначала выполнение условия нормировки для начальных состояний используя условия для начальных состояний из (6):

$$\sum_{q\in Q, r\in\overline{0,t}} P'_I(q,x^r) = \sum_{q\in Q} P'_I(q,\lambda) = \sum_{q\in Q} P_I(q) = 1.$$

Докажем теперь выполнение условий нормировки для переходов и конечных состояний. Для этого фиксируем $q\in Q$. Пусть $r\in\overline{0,t-1}$.

$$\begin{aligned} &P'_F(q,x^r) + \sum_{q'\in Q, b\in B} P'_T\left((q,x^r),b,\left(q',x^{r+1}\right)\right) + \sum_{q'\in Q, b\in B} P'_T\left((q,x^r),b,\left(q',x^r\right)\right) + \\ &+ \sum_{q'\in Q} P'_T\left((q,x^r),\lambda,\left(q',x^{r+1}\right)\right) = \\ &= \sum_{q'\in Q, b\in B} P_T\left(q,x_{r+1},b,q'\right) + \sum_{q'\in Q, b\in B} P_T\left(q,\lambda,b,q'\right) + \sum_{q'\in Q} P_T\left(q,x_{r+1},\lambda,q'\right) = 1. \end{aligned}$$

Пусть $r=t$. Тогда из трех сумм в предыдущей записи остается только средняя, поэтому запишем:

$$P'_F\left(q,x^t\right) + \sum_{q'\in Q, b\in B} P'_T\left(\left(q,x^t\right),b,\left(q',x^t\right)\right) = P_F(q) + \sum_{q'\in Q, b\in B} P_T\left(q,\lambda,b,q'\right) = 1.$$

При выполнении условий нормировки (7), единице будут равны левые части трех приведенных выше в данном доказательстве выражений, что доказывает выполнение условий (4). Утверждение доказано.

Доказательство утверждения 2. Пусть (q,x^u) было достижимо из (s,x^r), тогда существует путь $\theta'_{s,q} = \left((q_0,x^r),\hat{b}_1,\left(q_1,x^{r+i_1}\right),...,\left(q_{n-1},x^{r+i_{n-1}}\right),\hat{b}_n,\left(q_n,x^{r+i_n}\right)\right)$, имеющий ненулевую вероятность, где $q_0=s$, а $q_n=q$. Тогда по определению вероятности пути в автомате $\mathfrak{A}\left(\pi,x^t\right)$ и выражению (6) все вероятности переходов вида $P_T\left(q_{j-1},x_{r+i_j},\hat{b},q_j\right)>0$, если $i_{j+1}-i_j>0$, $P_T\left(q_{j-1},\lambda,\hat{b},q_j\right)>0$, если $i_{j+1}-i_j=0$. Это означает, что состояние q достижимо из s при помощи $(x_{r+1},...,x_{r+i_n})$.

Пусть теперь существует слово $\hat{a}\in\hat{A}$, получаемое из x^t путем вставки символов λ и путь $\theta_{s,q} = \left(q_0,\hat{a}_1,\hat{b}_1,q_1,...,q_{n-1},\hat{a}_n,\hat{b}_n,q_n\right)$ имеющий ненулевую вероятность. При этом $q_0=s$ и $q_n=q$. Тогда $P_T\left(q_{j-1},\hat{a}_j,\hat{b}_j,q_j\right)>0$, а следовательно соответствующие вероятности переходов автомата $\mathfrak{A}\left(\pi,x^t\right)$ также больше нуля, то есть состояние (q,x^u) достижимо из (s,x^r). Доказательство завершено.

Доказательство теоремы 1. Отметим, что аддитивность функции (5) следует из определения. σ-аддитивность будет следовать из сходимости соответствующего ряда. Таким образом, достаточно доказать, что

$$\sum_{y\in B*} P_\pi(y|x) = 1.$$

Пусть из состояния q достижимо состояние s при помощи $(x_1,...,x_r)$, из которого, в свою очередь, достижимо некоторое конечное состояние q'. Согласно утверждению 2, в вероятностном

автомате $\mathfrak{A}\left(\pi, x^t\right)$ из состояния (q, λ) достижимо состояние (s, x^r), из которого достижимо конечное состояние $\left(q', x^t\right)$. Согласно критерию порождения распределения вероятностей для автономного вероятностного автомата [11]:

$$\sum_{y \in B*} P_{\mathfrak{A}(\pi,x)}(y) = 1.$$

Согласно выражениям (5), (6) и утверждению 2.5, $P_{\mathfrak{A}(\pi,x)}(y) = P_\pi(y|x)$. Таким образом доказана корректность определения (5).

Пусть теперь функция $P_\pi(.|x)$ задает условное распределение вероятностей на B^*. Исходя из $P_{\mathfrak{A}(\pi,x)}(y) = P_\pi(y|x)$, отметим, что $P_{\mathfrak{A}(\pi,x)}(.)$ является распределением вероятностей. Тогда по критерию порождения распределения вероятностей для автономного вероятностного автомата [11] для любого начального состояния (q, λ), состояния (s, x^r), достижимого из (q, λ) найдется некоторое состояние $\left(q', x^t\right)$, достижимое из (s, x^r). Это по утверждению 2 означает, что из начального состояния q автомата π при помощи $(x_1, ..., x_r)$ достижимо состояние s, из которого при помощи $(x_{r+1}, ..., x_t)$ достижимо конечное состояние q'. Теорема доказана.

Доказательство теоремы 2. Если в УВ-автомате π все состояния достижимы, то $\chi_\pi = \pi$. Допустим, что множество состояний автомата π разбивается на два непересекающихся множества: $Q_\pi = Q_1 \cup Q_2$, где Q_1 – множество достижимых состояний, а Q_2 – недостижимых состояний. При этом $Q_2 \neq \emptyset$.

Рассмотрим автомат

$$\pi' = \left\langle Q'_\pi, \hat{A}, \hat{B}, T'_\pi, P'_{\pi,I}, P'_{\pi,F}, P'_{\pi,T} \right\rangle,$$

у которого:

- $Q'_\pi = Q_1$ – только достижимые состояния;
- $T'_\pi = T_\pi \backslash \left\{ (r, a, b, q) \,|r \in Q_2, a \in \hat{A}, b \in \hat{B}, q \in Q_\pi \right\}$ – переходы, начинающиеся только в достижимых состояниях;
- $P'_{\pi,I} = P_{\pi,I}\,|_{Q_1}$ – ограничение функции $P'_{\pi,I}$ на множестве достижимых состояний;
- $P'_{\pi,F} = P_{\pi,F}\,|_{Q_1}$ – ограничение функции $P'_{\pi,F}$ на множестве достижимых состояний;
- $P'_{\pi,T} = P_{\pi,T}\,\big|_{T'_\pi}$ – ограничение функции $P'_{\pi,T}$ на множестве переходов из достижимых состояний.

Все начальные состояния автомата π и только они являются начальными для автомата π'. Следовательно, условие нормировки из выражения (4) выполняется для начальных состояний автомата π'.

Поскольку любое достижимое состояние из Q_1 не имеет переходов в недостижимое состояние, то все оставшиеся суммы в оставшихся выражениях из (4) выполняются для автомата π' также как и для π, а следовательно π' является УКВ-автоматом. Очевидно, что все состояния в нем являются достижимыми. А из условия, что π является УВ-автоматом, следует, что из любого состояния множества Q_1 достижимо некоторое конечное состояние. При этом путь проходит только через достижимые состояния.

Таким образом, в автомате π' из любого состояния достижимо некоторое конечное состояние, то есть, согласно следствию из теоремы 3, π' является УВ-автоматом. Возьмем $\chi_\pi = \pi'$. Теорема доказана.

6. ВЫВОДЫ

В работе вводится общее понятие УКВ-автомата и важный частный случай УВ-автомата, доказываются критерий порождения УКВ-автоматом условного распределения вероятностей

$P_\pi(.|x)$ выходных слов для каждого входного слова (теорема 1) и эквивалентное определение УВ-автомата (следствие из теоремы 1). Используя данное следствие при проверке корректности УВ-автомата можно заменить вычисление всевозможных значений функции $P_\pi(.|.)$ на обход графа переходов автомата, что существенно упрощает проверку корректности.

Условие достижимости конечного состояния, обеспечивающее порождение автоматом условного распределения на множестве выходных слов конечной длины, не является существенным ограничением для применения рассматриваемой модели на практике. Действительно, если данное условие не выполняется, то найдется входное слово, которое будет порождать на выходе автомата "бесконечное"выходное слово, так как автомат никогда не попадет в конечное состояние, в котором возможен останов. Такая ситуация в рассматриваемой задаче моделирования ошибок распознавания смволов в словах конечной длины невозможна.

Введенное понятие эквивалентности УВ-автомато и доказанное наличие для каждого УВ-автомата эквивалентного ему УВ-автомата без недостижимых состояний (теорема 2) позволяет при их изучении не рассматривать случаи, связанные с наличием таких состояний.

При помощи рекуррентной формулы (8) вычисление значения функции $P_\pi(.|.)$ производится со сложностью $O\left(t \cdot v \cdot n^2\right)$. Следует отметить, что сложность алгоритма вычисления данной условной вероятности может быть уменьшена до $O(t \cdot v)$ при фиксации структуры переходов автомата.

Введенное в работе понятие УВ-автомата позволяет моделировать произвольный дискретный канал с ошибками в форме вставок, удалений и замещений символов, в том числе и для случая канала с памятью. С помощью множества состояний и определенной структуры переходов автомата можно задавать зависимость от различных факторов, например: значений входных символов с произвольной глубиной зависимости, числа входных символов, а также от ошибок, происходивших на предыдущих шагах работы алгоритма.

Таким образом, описанные свойства модели дискретного канала в форме УВ-автомата позволяют использовать ее для моделирования ошибок распознавания в условиях сложных зависимостей характеристик ошибок от различных факторов.

СПИСОК ЛИТЕРАТУРЫ

1. Brill E., Moore R. C. An Improved Error Model for Noisy Channel Spelling Correction // Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 2000, p. 286-293.

2. Kolak O., Resnik F. OCR Error Correction Using a Noisy Channel Model // Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, p. 55-62.

3. Ristad E. S., Yianilos P. N. Learning String Edit Distance // IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, v. 20(2), p. 522-532.

4. Szanser A. J. Automatic error-correction in natural languages // Proceedings of the 1969 conference on Computational linguistics, 1969, p. 1-8.

5. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР, 1965, т. 163(4), с 846-848.

6. Бухараев Р.Г. Основы теории вероятностных автоматов. – М.: Главная редакция физико-математической литературы, 1985.

7. Кудрявцев В. Б., Алешин С. В., Подколзин А. С. Введение в теорию автоматов. – М.: Главная редакция физико-математической литературы, 1985.

8. Трахтенброт Б. А., Барзинь Я. М. Конечные автоматы (поведение и синтез). – М.: Главная редакция физико-математической литературы, 1970.

9. Oncina J., Sebban M. Learning Stochastic Edit Distance: Application in Handwritten Character Recognition // Pattern Recognition, 2006, v. 39, p. 1575-1587.

10. Bernard M., Janodet J.-C., Sebban M. A Discriminative Model of Stochastic Edit Distance in the Form of Conditional Transducer // 8^{th} International Colloquium on Grammatical Inference, Japan, 2006, p. 240-252.

11. Dupont P., Denis F., Esposito Y. Links between Probabilistic Automata and Hidden Markov Models: probability distributions, learning models and induction algorithms // Pattern Recognition, 2005, v. 38, p. 1349-1371.

12. Vidal E., Thollard F., Higuera, de la, C., Casacuberta F., Carrasco R. C. Probabilistic Finite State Machines – Part I // IEEE Transactions on Pattern analysis and Machine Intelligence, 2005, v. 27(7), p. 1013-1025.