УДК 681.3.06

Н. М. Быков, к. т. н., доц.; Д. Е. Балховский; И. В. Кузьмин, д. т. н., проф.

АНАЛИЗ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК МОРФЕМ УКРАИНСКОГО ЯЗЫКА

Разработан алгоритм и программное обеспечение для определения статистических и переходных вероятностей морфем текста. Проведен теоретический анализ проблемы построения эффективной иерархической стратегии распознавания текста, а также предложена процедура построения оптимального дерева классификации образов текста.

Ключевые слова: *анализ статистических характеристик, морфемы, эффективная стратегия распознавания, скрытые марковские сети, оптимальная процедура классификации.*

Введение

Практика распознавания рукописных символов показывает, что использование только графической информации для их описания не позволяет получить удовлетворительных результатов с точки зрения скорости и надежности, поэтому возникает необходимость использования лингвистической информации, содержащейся в текстовом документе [1]. Естественно как такую информацию использовать контекстную, в качестве которой может служить лексическая и статистическая информация. Лексическую информацию удобно использовать тогда, когда элементами распознавания будут морфемы – минимальные содержательно разделимые части слова (например, приставка, корень, суффикс, окончание). Тогда можно использовать статистическую и лексическую информацию о морфемах для построения моделей слов текста в виде скрытых марковских сетей (СМС), что позволило бы применить известные алгоритмы распознавания на этих сетях. Кроме того процедура сегментации на морфемы будет выполняться значительно реже по сравнению с посимвольним распознаванием.

Постановка задачи

Авторами в работе [2] были разработаны программные средства для построения базы данных украинских морфем для обеспечения возможности использования морфологической информации в задаче распознавания текстового документа. В результате была построена база данных, которая содержит более 60 000 морфем украинского языка. Однако использование лишь одной базы данных без использования другой статистической информациии не позволит оптимизировать процесс распознавания текстов. Поэтому появляется необходимость в решении задачи определения статистических характеристик морфем в виде их статических и переходных вероятностей. Используя базу данных о статистических характеристиках, процесс распознавания рукописных и других текстов, написанных нетипизированными шрифтами, можно значительно ускорить за счет модульной иерархической архитектуры и аппарата скрытых марковских сетей (СМС). Такие сети позволяют после процедуры сегментации текста на морфемы и распознавания очередной морфемы в качестве альтернативы следующей морфемы выбрать морфемы с наибольшими переходными вероятностями из базы данных. То есть, исчезает необходимость в сравнении графического изображения морфемы (графемы) со всеми возможными эталонами. При этом принятие решений осуществляется выбором альтернативы с наибольшей суммарной вероятностью. Такой подход ввода, обработки и распознавания текстов повышает быстродействие и надежность всего процесса. Для реализации описанных идей в данной работе решаются задачи разработки эффективной стратегии распознавания текстового документа во время его ввода в компьютер и процедуры поиска оптимального

классификатора текстовых образов, а также разработки алгоритмов определения и анализа статистических характеристик морфем украинского языка с целью их использования на лексическом уровне распознавания.

Теоретический анализ проблемы построения эффективной иерархической стратегии распознавания текста

Любой текстовый документ можно рассматривать не только как графическое изображение, но и как некий носитель языковой информации, которая используется для ее передачи в той или другой коммуникативной системе [3]. С такой точки зрения графика текста опосредствованным образом отображает разные информационные уровни, свойственные коммуникативному акту: прагматический, семантический, синтаксический, лексический, морфологический, сигматический и аффектный [1]. Возникает вопрос: информацию какого уровня и в какой последовательности нужно использовать в автоматизированном процессе ввода и распознавания текстового документа, чтобы получить максимально возможную скорость и минимально возможные ошибки и стоимость. Для решения этого вопроса авторы в данной работе предлагают новую технологию электронизации текстовых документов, которая наряду с распознаванием графических образов использует частичное понимание текста. При этом процесс ввода рассматривается как процесс взаимодействия устройства ввода и языковый тезаурус компьютерной системы понимания текста. Во время сканирования изображения текста устройство ввода выделяет очередной признак графемы, относящийся к тому или иному информационному уровню языка, который используется системой для уменьшения энтропии о текстовой единице и сужения круга кандидатов на принятие решения (распознавание). В работе [3] авторы предложили формальную постановку задачи оптимизации процесса ввода и обработки текстового документа, которая рассматривает его в виде дерева классификации текстовых образов на разных информационных уровнях.

Оптимизация процесса распознавания образов текстового изображения осуществляется по информационному критерию эффективности

$$Э_{\text{Р}} = \frac{\text{I}_{\text{Р}}}{C_{\text{Р}}}, \tag{1}$$

предложенным в [3], где I_p – количество информации, которую получает система распознавания и понимания текста, определяется с учетом энтропийных свойств текстовых образов; $C_{\text{Р}}$ – стоимость системы;

$$C_{\text{Р}} = C_{\text{Х}} + C_{\text{К}}, \tag{2}$$

где $C_{\text{Х}}$ – сложность вычисления признакового описания образов; $C_{\text{К}}$ – сложность вычислений классификации образов.

Поскольку сложность C_p системы распознавания является адитивной суммой сложностей каждого из иерархических уровней распознавания, а информативность I_p является ненисходящей функцией вероятности правильного распознавания, то оптимальная стратегия является композицией алгоритмов распознавания, которые максимизируют отношение I_i / C_i на каждом из уровней. Последовательность композиции алгоритмов в оптимальной стратегии должна соответствовать последовательности размещения уровней дерева классификации, которые в свою очередь соответствуют информационным уровням текстового документа.

Решение проблемы надлежащего выбора коэффициента разветвления позволяет в значительной степени сузить круг поиска в оптимизационной процедуре поиска

оптимального дерева решений. В работе [4] показано, что минимизация суммарной ошибки классификации и времени классификации дает границы коэффициента разветвления B_r, которое выбирается во время построения оптимального дерева решений:

$$2 \le B_r \le 5. \tag{3}$$

Таким образом, установленные свойства дерева решений позволяют сузить диапазон поиска при решении задачи определения оптимального дерева классификации текстовых образов.

Решение задачи построения эффективной стратегии принятия решения в виде дерева классификации можно осуществить с помощью процедуры оптимизации "управляемого поиска вперед с возвращением" [4]. В этой процедуре критерий (2) руководит поиском такой структуры дерева решений среди всех возможных, в которой на каждом шагу поиска выбирается та конфигурация узлов, которая имеет наивысшее значение критерия. Для заданного узла Ω_i^h дерева процедура поиска выполняется в виде следующей последовательности шагов:

1. На основе выбранного признака $x^h \in X$ осуществляется одна из возможных разбивок $\pi^h \in \Pi$ узла Ω_i^h на подмножество узлов-потомков $\{\Omega_j\}, j = \overline{1,m}$. Признак x^h выбирается по "матрице различий" таким образом, чтобы коэффициент разветвления лежал в границах, определенных в (3). Здесь h – уровень (высота) дерева классификации, X – априорный алфавит признаков.

2. Вычисляется значение критерия (1) для полученной конфигурации узла.

3. Повторяя пункты 1 и 2, строят другие возможные разбивки и для них вычисляют значение критерия.

4. Определяют конфигурацию, для которой критерий имеет максимальное значение, и тем самым находят оптимальный набор признаков $\overline{B}_r$ для данного узла дерева и оптимальный шаг алгоритма классификации.

В качестве "матрицы различий" используют таблицу попарных различий графем текста w_i и w_j по всем признакам их описания из априорного алфавита признаков на основе выбранного в пространстве признаков расстояния d_{ij}.

Анализ статистических характеристик морфем

Представим графему слова на морфемном уровне дерева классификации в виде последовательности ***O*** векторов обсерваций:

$$O = \vec{o}_1, \vec{o}_2, ..., \vec{o}_L, \tag{4}$$

где $\vec{o}_l$ – вектор изображения морфемы.

Задача распознавания слов текста в таком случае может рассматриваться как вычисление максимума правдоподобности

$$\arg \max_i \{P(w_i / O)\}, \tag{5}$$

где w_i является *i*-тым словом словаря.

Согласно формуле Байеса

$$P(w_i / O) = \frac{P(O / w_i) P(w_i)}{P(O)} \tag{6}$$

наиболее вероятная графема слова в изображении текста определяется вероятностью. Прямая оценка совместной условной вероятности $P(\vec{o}_1, \vec{o}_2, ..., \vec{o}_L / w_i)$ из корпуса текста не

практикуется по причине астрономического количества возможных обсервированных последовательностей. В большинстве случаев задачу оценки плотности деления условных вероятностей $P(O/w_i)$ заменяют более простой проблемой оценки параметров Марковской модели генерации текста *M*. Эта модель представляет автомат с конечным количеством состояний, при установлении состояния *i* генерируется вектор изображения графемы $\vec{o}_l$ с вероятностью $b_i(\vec{o}_l)$. Кроме того, переход из состояния *i* в состояние *j* описывается вероятностью a_{ij}. Выбор наиболее вероятной графемы слова осуществляется путем нахождения наиболее правдоподобной последовательности состояний:

$$P(O/M) = \max_X \left\{ a_{x(0)x(1)} \prod_{l=1}^{L} b_x(\vec{o}_l) a_{x(l)x(l+1)} \right\}, \tag{7}$$

где $P_l(\vec{o}_l)$ – вероятность обсервации вектора изображения морфемы $\vec{o}_l$, $a_{x(l)x(l+1)}$ – вероятность перехода от графемы $\vec{o}_l$ к $\vec{o}_{l+1}$, X – множество состояний, которое воспроизводит модель. С учетом пространственной локализации состояний в Марковский модели текста авторами предложена модификация этой модели, которая заключается в дополнении требования (7) требованием:

$$\sum_{l=1}^{L} P_l(\vec{o}_{(l)}) = P(L), \tag{8}$$

где $P_l(\vec{o}_l)$ – математическое ожидание длины морфемы $\vec{o}_l$, $P(L)$ – математическое ожидание длины графемы слова w_i.

Выводы. Таким образом, для реализации на морфемном уровне алгоритма распознавания графемы слова текста (7) необходимо определить статические и переходные вероятности морфем в тексте и статистические характеристики длин морфем и слов (их изображений). В данной работе разработан алгоритм вычисления статических и переходных характеристик морфем на основе использования тестового корпуса текста и разработанной авторами в [5] базы данных морфем украинского языка. Результаты работы алгоритма фиксируются в виде двух матриц, в первой из которых фиксируются статические вероятности морфем, во второй – переходные вероятности. Таблица 1 демонстрирует вид второй матрицы.

В данной таблице приняты обозначения: m_1, m_2, .., m_N – морфемы украинского языка; N – количество морфем украинского языка; P (m_i/m_j) – вероятность перехода между i-ой и j-ой морфемами.

Таблица 1

Матрица переходных вероятностей морфем

Морфемы / Вероятности	Морфема 1	Морфема 2	...	...	...	...	Морфема N
Морфема 1	0	P (m_1/m_2)	...	...	...	...	P (m_1/m_N)
Морфема 2	P (m_2/m_1)	0	...	...	...	...	P (m_2/m_N)
...	...	...	0	...	...	...	...
...	...	...	...	0	...	...	...
...	...	...	...	...	0	...	...
...	...	...	...	...	...	0	...
Морфема N	P (m_N/m_1)	P (m_N/m_2)	...	...	...	...	0

Алгоритм создания матрицы переходных вероятностей морфем украинского языка состоит из следующих шагов:

1. Считывание базы данных (БД) морфем украинского языка.

2. Считывание массива слов из тестового корпуса текста, на котором будет построена БД вероятностей.

3. Запуск цикла от первой до последней морфемы (i = 1; i <= N), где N – количество морфем в БД.

4. Запуск внутреннего цикла от первой до последней морфемы (j = 1; j <= N), где N – количество морфем в БД.

5. Обнуляем счетчик (k) найденных i-ой и j-ой морфем, следующих одна за другой.

6. Запуск цикла от первого до последнего слова текста (w = 1; w <= M), где M – количество слов в тексте.

7. Поиск в w-ом слове i-ой и j-ой морфемы.

7. Если морфемы найдены и j-ая морфема следует за i-ой морфемой, инкрементируем счетчик k.

8. Возвращаемся к пункту 5 для перехода на следующее слово.

9. По завершении цикла 6 (пройдены все слова и посчитана сумма (счетчик k) найденных во всем массиве слов i-ой и j-ой морфем, следующих одна за другой) определяем вероятность перехода между i-ой и j-ой морфемами: $P(m_i/m_j) = k / N$.

10. Записываем определенную вероятность $P(m_i/m_j)$ в базу данных.

11. По завершении цикла 4 (пройдены все j-морфемы) возвращаемся к циклу 3.

12. По завершении цикла 3 (пройдены все i-морфемы) формируем отчет и выходим из программы.

Таблица 2

Пример определенных вероятностей

Морфемы / Вероятности	роз	піз	нав	ан	н	я
роз	0	0,0457	0,001	0,0255	0,0548	0,00023
піз	0,0652	0	0,0522	0,0453	0,0985	0
нав	0,0001	0	0	0,0781	0,0001	0,0012
ан	0,0001	0	0,00112	0	0,0268	0,0556
н	0	0	0	0,123	0	0,0434
я	0	0,002	0	0	0,0897	0

СПИСОК ЛИТЕРАТУРЫ

1. Пиотровский Р. Г. Текст машина, человек. — Ленинград: Наука", 1975. – 326 с.

2. Биков М. М. Використання інтелектуальних методів в розпізнаванні символів / М. М. Биков, Д. Є. Балховський, А. Раїмі // Інформаційні технології та комп'ютерна інженерія. – 2007. – № 2 (9). – С. 121 – 125.

3. Нова інформаційна технологія введення і оброблення текстових документів в автоматизованих інформаційно-пошукових системах // Автоматика-2008: доклади XV міжнародної конференції з автоматичного управління, 23 – 26 вересня 2008 р., – Одеса: ОНМА. – 992 с.

4. Биков М. М. Розробка ефективної стратегії прийняття рішень в комп'ютерних інтелектуальних системах / М. М. Биков // Вісник Хмельницького національного технічного університету. – 2005. – Ч.1. – Т. 2, № 2. – С. 22 – 30.

Быков Николай Максимович – к. т. н., доцент, профессор кафедры компьютерных систем управления, тел.: (0432)-598430, e-mail: nmbdean@ksu.vstu.vinnica.ua.

Балховский Дмитрий Евгеньевич – аспирант кафедры компьютерных систем управления, тел.: (0432)-598222, e-mail: vinbudya@yandex.ru.

Кузьмин Иван Васильевич – д. т. н., профессор, профессор кафедры компьютерных систем управления, тел.: (0432)-598222, e-mail: nmbdean@ksu.vstu.vinnica.ua.

Винницкий национальный технический университет.