УДК 681.14:621.01

ОБ ЭФФЕКТИВНОСТИ АВТОМАТИЧЕСКОГО КОДИРОВАНИЯ И ИСПРАВЛЕНИЯ ОШИБОК ПРИ ПОДГОТОВКЕ ДАННЫХ

В.В.Хабаров, Ю.Г.Косарев

Одно из серьёзных препятствий на пути широкого применения ЭВМ для автоматизации многих процессов - сложность подготовки информации. Как правило, на пользователя ЭВМ возлагается обязанность представлять информацию в виде, удобном для машины. Пользователь должен привычные для него термины естественного языка (точнее, его модификации, учитывающей специфику данной области) заменить специальными (обычно цифровыми) кодами.

Процесс такого кодирования заметно затрудняется как необходимостью обращаться к перекодировочным словарям, так и тем, что от пользователя требуется гораздо большая скрупулезность, чем та, к которой он привык при составлении документов, рассчитанных на человека. В то же время информация после подобного кодирования теряет свою наглядность, что затрудняет обнаружение ошибок "на глаз", как это обычно имеет место с текстами на естественном языке. Поэтому кодирование приходится повторять дважды, либо делать для контроля декодирование. Возможности автоматического обнаружения ошибок, а тем более их исправления, в данном случае ограничены из-за сложности, а нередко и невозможности построения формальных методов для обнаружения некоторых типичных видов ошибок (например, замена одного кода другим).

Важно также отметить, что этот трудоемкий и малопривлекательный процесс подготовки данных обычно приходится выполнять высококвалифицированным специалистам.

Естественно попытаться облегчить труд этих специалистов переложив его хотя бы частично на ЭВМ, т.е. предоставить человеку возможность описывать данные в привычных для него терминах естественного языка, а остальную часть работы: кодирование, обнаружение и исправление ошибок поручить машине.

Из общих соображений вполне понятно, что при словесном описании данных процесс кодирования должен ускориться, так как отпадает необходимость пользоваться перекодировочными словарями, становится ненужным повторное контрольное кодирование. Кроме того, что весьма важно, из-за большей наглядности информации можно ожидать уменьшения числа формально не обнаруживаемых ошибок, а увеличение избыточности позволяет надеяться на увеличение доли автоматически исправляемых ошибок.

Наряду с этим может заметно увеличиться объем записи, что повлечет за собой рост времени собственно записи у кодировщика и времени перфорации.

Таков общеизвестный качественный характер ожидаемых изменений при переходе к естественному языку (словесному кодированию).

Вполне понятно, что вывод о практической целесообразности перехода к словесному кодированию можно сделать лишь на основании к о л и ч е с т в е н н ы х оценок и при непременном условии - создании простых и эффективных программных средств для реализации автоматического кодирования, обнаружения и исправления ошибок.

Далее описывается эксперимент, который позволяет получить требуемые количественные оценки для одной из областей применения - автоматизации проектирования маршрутной технологии изготовления деталей машин,- и предлагается методика автоматического кодирования и исправления ошибок при подготовке данных.

При этом за основу взята одна из реальных систем автоматизации маршрутной технологии, применяемых при проектировании промышленных предприятий.

Для проектирования маршрутной технологии, т.е. состава и размещения оборудования, наиболее опасны ошибки, которые могут привести к неправильному выбору вида оборудования (например, типа станка) либо его марки (т.е. параметров станка). Иными словами, опасны искажения при кодировании и больше искажения значений числовых параметров. Так как последние

можно относительно просто обнаружить, сравнивая их с габаритами детали, то основное внимание в данной работе уделяется процессу кодирования.

I. Методика автоматического кодирования и исправления ошибок. Решалась следующая задача. Задан словарь, каждому слову которого поставлен в соответствие код. Требуется найти для каждого слова входного текста соответствующее слово из словаря и заменить его кодом.

При отсутствии искажений задача не вызывает каких-либо затруднений. При искажениях нужно найти слово, наиболее близкое к данному, т.е. возникает типичная задача распознавания образов [I]. Методы решения подобных задач обычно опираются на трудоемкую процедуру сравнения данного слова со всеми эталонными словами (словами из словаря), поэтому применялся более простой метод, который теоретически не всегда приводит к успеху, но зато весьма прост в реализации.

Суть этого метода [2] заключается в следующем. Анализируется слово буква за буквой до тех пор, пока не встретится буква, которой данное слово отличается от всех других слов (рис.I). После этого слову присваивается код, а затем проверяется окончание, при совпадении осуществляется переход к следующему слову.

I. Полоса
2. Проволока
3. Пружина
4. Пруток

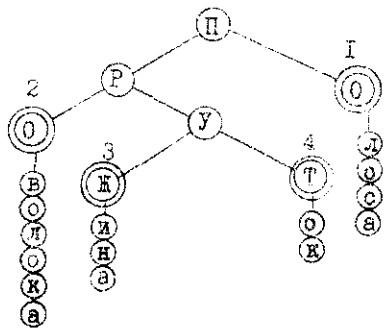

Рис. I

Буквы, которые отличают данное слово от всех других слов при анализе с начала слова (I), определяют первую базу, при анализе с конца слова (П) – вторую базу. (На рис. I первые базы подчеркнуты снизу, вторые – сверху.)

Если при анализе слова окончания не совпадают, то слово заносится в список исправлений.

При ошибке в базе анализ слова повторяется по другой базе. Если ошибка при определении этой базы не обнаружена, то слову присваивается код и оно заносится в список исправлений. При ошибке и во второй базе слово заносится в список неисправленных ошибок. Оба списка печатаются в удобной для пользователя форме. Каждой ошибке присвоено имя. В списке исправлений каждое слово с ошибкой сопровождается тем словом, на которое оно исправлено, т.е. пользователю достаточно для контроля правильности исправлений бегло просмотреть этот список.

В качестве языка программирования применяется аппарат формальных грамматик, аналогичный тому, который был предложен И.В.Вельбицким [3].

Важная особенность предлагаемой методики – простота ее использования. От пользователя требуется лишь словарь с указанием кодов. Составление соответствующих программных средств (R –таблиц) для кодирования, обнаружения и исправления ошибок выполняется автоматически.

2. Эксперимент.

2.I. Структура данных. Информация о детали обычно представлена в виде двух разделов: описания общих параметров детали и описания параметров ее элементов (рис.2). Технолог при кодировании использует словарь, который содержит наименования материалов, виды заготовок, названия элементов поверхностей, операции обработки, обозначения параметров.

Описание детали и каждого ее параметра состоит из нескольких уровней записи: символа (u_0), слова (u_1), предложения (u_3), фразы (u_4). При цифровом кодировании код соответствует предложению, группа кодов – фразе, т.е. уровни u_0 и u_1 отсутствуют.

Фразы состоят из последовательности предложений. Первое – определяет название детали или ее элемента, последующие – наименования параметров и их значения. Предложение может состоять из одного слова. Словами языка служат слова из словаря, их со –

№ черт.	издел.	матер.	К	Q	ТН84100I7, колесо зубчатое,
84100I7	I	0,8	4	I4	сталь эи-474, К4, Q I4
загот.	ϕ,B	L	Н	S	пруток, Д19.8, L I4.5,
02	I9,8	I4,5	0	0	Т5, Ч5
№	вод	параметры		Т7	
I	0I	I5	I3	I	55	торец, Д15, L1
2	02	I5	6	0	55	цилиндр, Д15, L6
3	26	I9,8	I5	2,4	55	уступ, Д19.8, L2.4
4	02	I9,8	5	0	56	цилиндр, Д19.8, L5, Ч6
5	26	I9,8	I5	2,4	55	уступ, Д19.8, L2.4
6	02	I5	3,5	0	55	цилиндр, Д15, L3.5
7	0I	I5	7	4	55	торец, Д15, L4
8	09	7	5	0	27	отверстие, Д7, L5, Т2, Ч7
9	I0	I3	7	3	55	дно отверстия, Д13, L3
I0	09	I3	9,5	0	27	отверстие, Д13, L9.5, Т2, Ч7
II	42	I9,8	0,3	64	27	зубья цилиндрические, Д19.8, М0.3, z64, L5, Ч7

Рис.2. Зубчатое колесо. Примеры кодирования

кращения, буквы, обозначающие параметры, а также числовые значения параметров. В качестве символов используются все русские буквы, некоторые латинские, а также цифры и разделители.

2.2. Описание эксперимента. За основу словесного кодирования был взят полный вариант языка, который представлен примером на рис.2. Кроме того, был рассмотрен вариант языка (Ш), в котором допускались сокращения слов.

В эксперименте на этапе кодирования информации участвовали две равноценные по квалификации группы технологов, прошедших предварительное обучение. Квалификационная структура каждой группы отражала реальный состав групп проектного производства. Кодировались две партии чертежей (партии *А* и *В*) по 800 чертежей в каждой (табл.I). Чертежи подбирались с учетом практической равноценности деталей по их технологической сложности.

Данные перфорировались группой операторов на телеграфных аппаратах в цифровом коде ЭВМ "Минск-22" и в коде "М-2". Перфолента проверялась методом дубль-перфорации с последующим сравнением на контрольно-считывающем устройстве.

В ходе эксперимента применялась следующая методика обнаружения и исправления ошибок:

- при цифровом кодировании (вариант I) вся информация повторно кодировалась (без записи), и результат сравнивался с предыдущим (обнаруженные в тексте ошибки тут же исправлялись);
- при словесном кодировании (вариант П и Ш) непосредственно проверялось соответствие между элементами чертежа и их записью (ошибки также исправлялись сразу же после обнаружения);
- при перфорации (для всех вариантов) применялась повторная перфорация, сверка и печать исправлений.

Общая продолжительность эксперимента с учетом времени на обучение составила 7 месяцев.

Для оценки вариантов кодирования измерялись следующие параметры:

- трудоемкость подготовки данных по видам работ;
- объем информации;
- количество, вид и тип ошибок.

Рассматривались три вида ошибок: пропуск (x), дополнение (y) и замена (z) записи, которые свойственны каждому уровню записи. Эти ошибки, в зависимости от возможности их обнаружения и исправления, классифицировались по следующим типам:

Таблица I

Обработка двух партий первичных документов А и В (А=В=800)

Этапы подготовки данных	Исполнители	Количество человек в группе	Средний оклад, руб.	Средний коэффициент квалификации	Количество обрабатываемых исполнителем документов по вариантам языка I	Количество обрабатываемых исполнителем документов по вариантам языка II	Количество обрабатываемых исполнителем документов по вариантам языка III
Этап I	I группа Инженеры-технологи	4	I40	I,4	A/2	B	-
Заполнение первичных документов	2 группа Инженеры-технологи	4	I40	I,4	B/2	-	A
	Всего первичных документов				800	800	800
Этап 2 Перфорация данных	Операторы	4	80	0,8	200	200	200

A – автоматически исправляемые;

B – автоматически обнаруживаемые, но не исправляемые;

C – семантические (синтаксически не обнаруживаемые).

Методика обнаружения и исправления ошибок описана выше (п.I).

3. О б с у ж д е н и е р е з у л ь т а т о в э к с п е р и м е н т а.

3.I. Сравнение цифрового и словесного кодирования. Переход от цифрового кодирования (вариант I) к словесному (вариант II) привел (табл.2-4) к следующему:

- трудозатраты инженеров (этап I) сократились примерно на 40%;
- трудозатраты перфорации (этап II) возросли примерно вдвое;
- общие трудозатраты уменьшились мало (примерно на I0%).

Однако если учесть качество труда, то выигрыш получится вполне ощутимым. Даже при оценке качества труда по различию в зарплате выигрыш в стоимости получается около 20%. Если учесть различие во времени подготовки инженера (5 лет) и перфоратор-щицы (0,5 года), то выигрыш составит около 35% (из 40% возможных).

Указанное сокращение трудозатрат на кодирование (на 40%) произошло из-за применения кодов, непосредственно отражающих семантику терминов, которые они обозначают (эффект наглядности).

При этом около 3/8 этого выигрыша проистекает от упрощения самого процесса кодирования, так как отпала необходимость обращения к словарю, хотя и увеличился (в 2,7) объем записи. Остальные 5/8 выигрыша приходятся на контроль. Здесь две основные причины: исключение обращений к словарю и уменьшение числа исправлений в тексте.

Число ошибок при кодировании уменьшилось примерно в 2,7 раза, что особенно важно, более чем на порядок сократилось число ошибок, обнаружение которых формальными методами особенно сложно. Более трети всех обнаруженных ошибок исправляются с помощью описанной выше методики.

Число ошибок перфорации возросло примерно в 2,I раза при росте объема данных в 2,7. Непропорциональность в увеличении числа ошибок подтверждает известный факт, что осмысленный текст печатать проще. Все ошибки перфорации, что весьма важно, ока-

Таблица 2

Среднестатистические оценки на I первичный документ

Вариант языка	Этапы подготовки данных	Объем знаков	Трудоемкость обработки, чел/мин			Распредел. трудоемк. по этапам, %	Производит.труда в сравнении с вариантом I		Количество обнаружен. ошибок	Распредел. ошибок по этапам, %
			подготовка	контроль	общая трудоемкость		ручной контроль и исправл.ош.	обнаружен. и исправл. ошибок на ЭВМ		
I	Этап I	I40	I5,3	I2,8	28,I	78	I00	I00	0,I73	54
	Этап 2		2,I	5,7	7,8	22	I00	I00	0,I46	46
	Итого		I7,4	I8,5	35,9	I00	I00	I00	0,3I9	I00
II	Этап I	380	II,2	5,7	I6,9	52	I66	25I	0,064	I7
	Этап 2		4,9	I0,7	I5,6	48	50	I60	0,304	83
	Итого		I6,I	I6,4	32,5	I00	II0	223	0,368	I00
III	Этап I	270	I0,8	6,6	I7,4	53	I6I	260	0,I00	24
	Этап 2		4,0	II,3	I5,3	47	5I	I95	0,3I8	76
	Итого		I4,8	I7,9	32,7	I00	II0	243	0,4I8	I00

Таблица 3

Стоимость обработки документов

Вариант языка	Этапы подготовки данных	Стоимость обработки I первичного документа, руб.			Стоимость в сравнении с вариантом I, %	
		подготовка	контроль и исправлен.ошиб.	общая стоимость	ручной контроль и исправлен.ошиб.	обнаружение и исправление ошибок на ЭВМ
I	Этап I	0,2I4	0,205	0,4I9	I00,0	I00,0
	Этап 2	0,0I7	0,050	0,067	I00,0	I00,0
	Итого	0,23I	0,255	0,486	I00,0	I00,0
II	Этап I	0,I57	0,09I	0,248	59,2	37,5
	Этап 2	0,039	0,096	0,I35	20I,5	58,2
	Итого	0,I96	0,I87	0,383	78,8	40,3
III	Этап I	0,I5I	0,I06	0,257	6I,3	-
	Этап 2	0,032	0,I02	0,I34	200,0	-
	Итого	0,I83	0,208	0,39I	80,5	-

Таблица 4

Количество ошибок по вариантам языка

Виды ошибок	I								II							
	Этап I				Этап 2				Этап I				Этап 2			
	Всего	в том числе			Всего	в том числе			Всего	в том числе			Всего	в том числе		
		Обнаружимые	Неправильных			Обнаружимые	Неправильных			Обнаружимые	Неправильных			Обнаружимые	Неправильных	
			по базе I	по базе 2			по базе I	по базе 2			по базе I	по базе 2			по базе I	по базе 2
x_0 y_0 z_0	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	4 0 II	4 0 II	3 0 7	I 0 4	4 0 II	37 54 I07	I6 34 68	I8 20 38	34 54 I06
x_1 y_1 z_1	- - -	- - -	- - -	- - -	- - -	- - -	- - -	- - -	7 I I6	7 I 23	0 0 0	0 0 0	I 0 0	II 8 20	0 0 0	0 0 0	0 0 0
x_2 y_2 z_2	I2 7 I3	0 0 0	0 0 0	0 0 0	20 I9 62	20 I9 62	0 0 0	0 0 0	4 0 0	4 0 0	0 0 0	0 0 0	4 0 0	5 I 0	0 0 0	0 0 0	0 0 0
x_3 y_3 z_3	29 48 0	0 0 0	0 0 0	0 0 0	I6 0 0	I6 0 0	0 0 0	0 0 0	I 0 0	0 0 0	0 0 0	0 0 0	I 0 0	0 0 0	0 0 0	0 0 0	0 0 0
Всего по уровням	I38	32	0	0	II7	II7	0	0	5I	43	I0	5	I5	243	II8	76	I94
Всего по варианту	255								294								

зались формально обнаруживаемыми. Около 80% из них исправляются с помощью простейшей методики, описанной выше. Эффект "осмысленности" текста в данном эксперименте, по-видимому, проявился не полностью, так как перфораторщицы, участвующие в эксперименте, обладали в основном опытом работы с числовыми данными.

3.2. О возможности отказа от ручного контроля. Как пока - зал эксперимент, на 400 деталей встретилось всего 8 ошибок, которые формально не обнаруживаются методами, описанными выше. В семи случаях кодировщик неправильно назвал вид поверхности и в одном пропустил фразу. Допустимо ли такое число ошибок - вопрос дискуссионный. Однако одно из серьёзных последствий этих ошибок - неправильный выбор парка оборудования - может быть в значительной мере устранено при анализе малозагруженного оборудования на следующих этапах проектирования.

В целом вопрос об отказе от ручного контроля, который при словесном кодировании занимает около I/3 времени кодировщика, требует дополнительного исследования.

Все ошибки перфорации оказались обнаруживаемыми и в подавляющем большинстве исправимыми. Специальное исследование ошибок перфорации [4] также подтверждает данный вывод. Эти результаты дают основание сделать вывод о возможности для определенных групп перфораторщиц отказаться от повторной перфорации. В согласии с данными эксперимента (табл. 2,3), это более чем втрое сократит трудозатраты на перфорацию при словесном кодировании, и они станут меньше (примерно в I,6 раза), чем при цифровом кодировании.

3.3. Сравнение вариантов словесного кодирования. Полная форма словесного кодирования дает увеличение количества символов в 2,7 раза. Возникает вопрос, нельзя ли уменьшить это соотношение путем перехода к сокращенным словам, применить такие сокращения, которые не привели бы к потере наглядности? Вопрос однозначно не решается, поскольку при сокращении объема уменьшается и избыточность, т.е. возникает задача на оптимум, основным критерием которой является возможность автоматического исправления ошибок.

Для рассмотренного сокращенного варианта языка (Ш) (табл.2) эксперимент не выявил каких-либо существенных преимуществ перед вариантом (П) по производительности. В то же время

число ошибок в этом варианте больше, чем у варианта П. Особенно плохо то, что из-за уменьшения избыточности произошло увеличение числа неисправляемых ошибок [4], а это не позволяет применить для этого варианта режим автоматического исправления ошибок и отказаться от ручного контроля.

Таким образом, введение сокращенного языка требует дополнительных исследований. Поэтому в настоящее время для практических целей обоснованным является применение полного варианта языка.

Предлагаемая методика подготовки нечисловой информации для ЭВМ даже в простейшем ее виде, благодаря применению привычного для пользователя языка и переложению на машину процессов кодирования, обнаружения и исправления ошибок, сокращает примерно вдвое трудозатраты по сравнению с ручным кодированием и контролем.

Литература

1. ЗАГОРУЙКО Н.Г. Методы распознавания и их применение. М., "Сов.Радио", 1972.

2. КОСАРЕВ Ю.Г., КОНСТАНТИНОВ В.И., НУРИЕВ Р.М. Метод эффективной обработки текстовой информации (*TR* -грамматики). Отчет Института математики СО АН СССР, 1974.

3. ВЕЛЬБИЦКИЙ И.В. Метаязык *R* -грамматик. -"Кибернетика", 1973, № 3, с.47-63.

4. ПИЛИПОВИЧ Ю.В., КОСАРЕВ Ю.Г. Статистика ошибок перфорации. Отчет института математики СО АН СССР, 1974.

Поступила в ред.-изд.отд.
19 декабря 1974 года