На правах рукописи

Трояновская Ольга Вадимовна

АВТОМАТИЗИРОВАННЫЕ СИСТЕМЫ УПРАВЛЕНИЯ И ОБРАБОТКИ ИНФОРМАЦИИ ДЛЯ АРХИВОВ МЕДИЦИНСКИХ ДОКУМЕНТОВ

Специальность: 05.13.09 — управление в биологических и медицинских системах (включая применение ВТ)

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург — 1999

Работа выполнена в Санкт-Петербургском государственном электротехническом университете (ЛЭТИ).

Научный руководитель -
кандидат технических наук, доцент Манило Л. А.

Официальные оппоненты:
доктор технических наук, профессор Фомин Б. Ф.
кандидат технических наук Шаповалов В. В.

Ведущая организация -
Санкт-Петербургская медицинская академия последипломного образования

Защита состоится "15" декабря 1999 г. в 10 часов на заседании диссертационного совета Д063.36.09 Санкт-Петербургского государственного электротехнического университета (ЛЭТИ) по адресу: 197376, Санкт-Петербург, ул. Проф. Попова, 5.

С диссертацией можно ознакомиться в библиотеке университета.

Автореферат разослан "2" ноября 1999 г.

Ученый секретарь
диссертационного совета Юлдашев З.М.

Р. с 218, D

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В настоящее время в обеспечении эффективной работы архивов медицинских документов все большую роль играет автоматизированная обработка больших массивов информации. Основными требованиями, предъявляемыми к автоматизированным системам управления и обработки информации (АСУиО), являются следующие:

1. Быстрая подготовка и адекватное представление входной медицинской информации.
2. Оперативное предоставление полных и точных ответов на запросы пользователей.
3. Возможность проводить централизованную обработку данных, принадлежащих медицинским архивам различных лечебных учреждений.

Вопросы, относящиеся к электронной форме представления медицинской информации о пациенте, способам автоматизированной обработки медицинских сведений на естественном, в частности, русском языке, а также методам быстрого поиска данных, обеспечивающих наиболее полный и точный ответ на запрос пользователя, не проработаны в полной мере.

Автоматизированные системы, эксплуатирующиеся сегодня в российских архивах медицинских документов, решают узкоспециализированные задачи и работают с неполной или искаженной информацией об объекте исследования, представленной набором кодов из стандартных классификаторов. Принципы, на которых основана подготовка данных для таких систем, не обеспечивают возможности эффективного ввода данных, требуют больших затрат ручного труда медицинских работников.

Важность адекватного компьютеризированного представления медицинской информации о пациенте обусловлена переходом медицинского обеспечения населения на принципы страховой медицины и подготовкой лечебным учреждением возрастающего потока статистической отчетности. Электронные записи пациентов должны соответствовать определенным стандартам для создания национальных регистров, а также интеграции с другими автоматизированными системами в России и за рубежом.

В связи с вышеизложенным представляется актуальным создание удовлетворяющих современным требованиям АСУиО для архивов медицинских документов.

Цель диссертации: разработка принципов и методов автоматизированного анализа и классификации естественно-языковых (ЕЯ) медицинских данных о пациенте, а также реализация их в автоматизированной системе обработки документов медицинского архива.

Задачи диссертации:

1. Исследовать методы и принципы обработки информации в естественно-языковой форме как в автоматизированных медицинских архивах, так и в других областях применения.
2. Синтезировать лингвистическую модель для анализа медицинских сведений на естественном языке с последующей их формализацией.
3. Построить модель интерпретации для работы с абстрактным представлением медицинских сведений о пациенте.
4. Разработать язык отображения естественно-языковых сведений о пациенте на модель интерпретации.
5. Провести тестирование моделей на объекте исследования.
6. Создать формальную систему, позволяющую оперировать абстрактными представлениями медицинских сведений о пациентах, и проверить применимость системы для решения задач классификации последних по запросам пользователей.

Методы исследований. Теоретические и прикладные разделы диссертации разработаны с применением теории формальных грамматик, теории распознавания образов (РО), теории операций, теории множеств, теории матриц, операций математической логики, элементов математической статистики.

Экспериментальные исследования проводились в Архиве военно-медицинских документов. В качестве объекта исследования использованы архивные материалы в виде историй болезни 4-х тысяч раненых и 6-ти тысяч больных военнослужащих, принимавших участие в локальных конфликтах.

Новые научные результаты.

1. Синтезирована лингвистическая модель представления медицинских сведений о пациенте на русском языке, позволяющая проводить формализацию персональной медицинской информации в терминах тезауруса предметной области.
2. Построена модель интерпретации для описания формализованных медицинских сведений о пациенте. Модель имеет 5 разновид-

ностей по числу лингвистических групп входного текста на профессиональном медицинском языке.

3. Предложен и реализован в автоматизированной системе метод лингвистической обработки персональных медицинских документов, отображающий ЕЯ информацию в формализованное представление.
4. Построена алгебра цепочек, которая позволяет оперировать формализованными представлениями медицинских сведений и включает восемь операций. В основу алгебры положена комбинация логических и структурных методов РО.
5. Разработан классификационный алгоритм, определяющий сходство формализованных медицинских сведений пациента и запроса пользователя на основе алгебры цепочек.

Основные положения, выносимые на защиту.

1. Модель предметной области автоматизированных систем для медицинских архивов должна быть представлена гибридным способом. Основные составляющие модели: тезаурус понятий предметной области, лингвистическая модель естественно-языковых медицинских сведений о пациенте и модель интерпретации для формализации этих сведений.
2. Для корректной автоматизированной обработки естественно-языковых медицинских сведений о пациенте с последующим преобразованием в формализованное представление используют знания семантико-синтаксической структуры текста, которая описана лингвистической моделью.
3. Обеспечить высокое качество автоматической классификации формализованных описаний пациентов в соответствии с запросами пользователей можно при учете не только синтаксиса, но и семантики формализованных представлений, что достигается применением структурных и логических методов распознавания образов. Реализацию классификации необходимо осуществлять в алгоритме, использующем операции специально разработанной алгебры цепочек.

Практическая ценность работы.

1. Предложен метод полуавтоматического приобретения знаний из медицинских текстов для построения тезауруса предметной области в сфере персональной медицинской информации.
2. Разработан алгоритм кодирования, который в ходе тестирования на 896-ти медицинских документах пациентов показал высокое качество автоматической формализации данных, исключающей субъективный фактор при обработке медицинской информации.

3. Предложена архитектура автоматизированной системы, выполняющей кодирование медицинских сведений о пациенте и их классификацию в соответствии с запросами пользователей, для работы в медицинских архивах.
4. Разработанный классификационный алгоритм протестирован на 896-ти реальных медицинских документах. Получены результаты: отклик системы равен 0.99, точность поиска - 1.00, что превышает показатели зарубежных аналогов.

Реализация результатов работы. Предложенные методы и алгоритмы реализованы в прототипе автоматизированной системы обработки данных, который эксплуатируется в Архиве военно-медицинских документов.

Апробация работы. Основные результаты и положения диссертационной работы докладывались и обсуждались на: Всероссийской научной конференции "Медицинская информатика накануне 21 века" (г. Санкт-Петербург, 1997 г.); Международном семинаре "Biomedical Engineering & Medical Informatics'97" (г.Гливице, Польша, 1997 г.); научно-технической конференции "Диагностика, информатика, метрология, экология, безопасность – 98" (г.Санкт-Петербург, 1998 г.); Международной конференции по мягким вычислениям и измерениям-99 (г.Санкт-Петербург, 1999 г.); 2-й Международной технической конференции "Медико-экологические информационные технологии - 99" (г. Курск, 1999 г.); научно-технических конференциях профессорско-преподавательского состава СПбГЭТУ (1998-1999 гг.).

Публикации. По теме диссертации опубликовано 8 печатных работ, из них 4 статьи и 4 доклада на конференциях и семинарах.

Структура и объем диссертации. Диссертационная работа состоит из введения, четырех глав с выводами, заключения, списка литературы, включающего 100 наименований, трех приложений. Основная часть работы изложена на 148 машинописных страницах. Работа содержит 11 рисунков и 10 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, сформулированы требования, предъявляемые к автоматизированным системам управления и обработки информации для медицинских архивов, определены цель и задачи диссертационной работы, основные положения, выносимые на защиту, изложено краткое содержание работы.

В первой главе рассмотрены основные проблемы обработки ЕЯ сведений о пациентах, которые стоят перед медицинскими учреждениями. В работе российских медицинских архивов существует высокая загрузка врачей рутинной подготовкой различных документов; выходная информация в виде отчетов, сводок и т.д. имеет низкое качество и длительные сроки исполнения. Проведен обзор методов решения подобных проблем в зарубежных госпиталях. Следование стандартам, автоматическая обработка данных и переход к компьютеризированным историям болезни дают положительные результаты. Вместе с тем отмечается, что автоматическое кодирование по стандартным классификаторам приводит к значительной потере и искажению обрабатываемой информации о пациенте. Альтернативой служит ЕЯ представление информации, но такой способ не обеспечивает полноту выборки информации по запросу. Описаны разнообразные модели медицинской информации, используемые в западных медицинских учреждениях и ориентированные на английский, немецкий и французский языки. Проведен анализ существующих систем и выделены основные функции, которые должны ими выполняться: а) способность к логическому выводу, б) обеспечение поддержки принятия решений, в) понимание ЕЯ информации, г) способность к обучению. Сформулированы требования к проектированию автоматизированных систем. Перечислены результаты, получаемые от внедрения интеллектуальных систем в медицинские архивы.

Во второй главе рассмотрены формы и методы представления знаний для выбора модели предметной области. Обоснована необходимость использования нескольких формализмов с целью построения гибридной модели, включающей тезаурус понятий предметной области, лингвистическую модель ЕЯ медицинских сведений, модель интерпретации для формализации этих сведений, механизм отображения лингвистической модели на модель интерпретации. Основное предназначение тезауруса - обеспечение набора понятий для гомоморфного отображения ЯЗ сведений о пациенте или запросе пользователя в формализованное представление. Описаны две составляющие тезауруса - лексико-морфологический словарь медицинских понятий, включающий в себя термины, их синонимы, родственные слова, варианты написания и сокращений, и семантическая сеть (СС), в которой медицинские понятия представлены узлами, а отношения, их связывающие, - дугами. Это позволило каждому ЕЯ

медицинскому понятию словаря поставить в соответствие его формализованный эквивалент - узел СС. Определено четыре класса отношений на сети: родовидовые, функциональные, грамматические, причинно-следственные. Количество узлов сети превышает 1000, число статей - 1500. Тематические деревья образованы дугами с родовидовыми отношениями, соединяющими семантически родственные узлы. В работе сформировано 13 таких деревьев.

Проектирование модели интерпретации выполнено на базе синтаксического подхода РО. Модель, предназначенная для описания на формализованном языке ЕЯ сведений о пациенте и поисковых запросов, представлена в виде:

$$\prod_{i=1}^{N} W_i^{+},\ W_{2i-1} = U_m,\ W_{2i} = C_k, \qquad (1)$$

где $\prod_{i=1}^{N}$ - конкатенация N элементов W_i, N - нечетное число; U_m, C_k - условные обозначения (у. о.) соответственно m-го медицинского понятия и k - ой взаимосвязи между понятиями в цепочке (см. табл.); знак + означает присутствие в цепочке, по крайней мере, одного элемента; $C_k \in \{$"@", ".", ",", "&", "+", "-", "*", ">"$\}$.

Перечень взаимосвязей между медицинскими понятиями

№ п/п	Название взаимосвязи	Условное обозначение	Тип взаимосвязи
1	Принадлежность	@	мягкая
2	Равноправие	.	мягкая
3	Перечисление	,	мягкая
4	Соединение	&	жесткая
5	Дополнение 1-го порядка с включением	+	жесткая
6	Дополнение 1-го порядка без включения	-	жесткая
7	Дополнение 2-го порядка	*	жесткая
8	Причинность	>	жесткая

Третья глава посвящена синтезу лингвистической модели, описывающей ЕЯ сведения. Для построения модели проанализирована обучающая выборка с текстами из 5-ти тысяч историй болезни. Сделано заключение, что тексты ограничены по тематике и

используемым лингвистическим структурам, т. е. язык медицинских документов (внешний язык) требует построения специализированной лингвистической модели.

В работе обоснована целесообразность построения модели на основе индексной грамматики, определены критерии построения. Описан метод полуавтоматического создания алфавита грамматики внешнего языка, который заключается в построении полного конечного множества используемых терминальных и нетерминальных элементов. Для этого в текстах на внешнем языке были выделены 5 групп сходных по своей форме и законченных смысловых сегментов текста, им поставлены в соответствие группы описания: 1) "Ранения, травмы, ожоги/отморожения" (РТО), 2) "Заболевания", 3) "Осложнения", 4) "Последствия РТО", 5) "Лечение". Внутри групп выделена устойчивая подгруппа «Локализация», описывающая анатомическую локализацию организма, которая может входить в состав любой группы, а также другие повторяющиеся подгруппы. В каждой группе было выделено одно главное - опорное слово (ОС) и его характеристики - дополняющие и описывающие его слова.

Все лексические единицы (ЛЕ) сгруппированы по определенным критериям в 17 категорий: ОС и характеристики группы «РТО»; ОС и характеристики подгруппы «Локализация»; функциональные слова, выражающие отношения между ЛЕ, (*по, вследствие, в<результате|виде|стадии|форме>, с|со, без, после, по <типу|линии>, осложненный;* где в < > через | - варианты контекста) и т. д. Каждой группе, подгруппе описания и категории присвоены имена, составляющие нетерминальный словарь. Терминальный словарь включает все ЛЕ с частотой встречаемости выше заданного порога.

Подробно описан разработанный метод ручного вывода грамматики, приведены правила подстановки. Предложен метод лингвистической обработки сведений о пациенте с использованием построенной модели. На первой стадии каждой ЛЕ присваивают одну из 17-ти категорий. Вторая стадия состоит из 6-ти этапов и оптимизирована по времени обработки путем эвристического подбора как группы описания анализируемого сегмента текста, так и набора правил грамматики. В случае неверного синтаксиса документ изымается из обработки. Проверенные сегменты поступают на стадию семантической интерпретации для их формализации.

Семантика внешнего языка L описана в виде $S=\varphi(L)$, где S - множество смыслов, φ - интерпретирующее отображение L в S. Элементы множества S - цепочки формальной порождающей грамматики внутреннего языка. Для формализации текста отображают: ЕЯ медицинское понятие в у. о. узла СС; функциональное слово между медицинскими понятиями в у. о. взаимосвязи. Языку L поставлена в соответствие денотационная семантика на основе интерпретации $I=(D,I_m,I_f)$, где I_m и I_f - логические функции, I_m соотносит каждое ЕЯ понятие с понятием тезауруса, I_f - каждое функциональное слово с взаимосвязью при учете контекста, D - область интерпретации - множество грамматически правильных цепочек. Алгоритм работы функции I_m сведен к поиску понятия тезауруса, описанного таким же набором слов, что и анализируемое понятие в тексте, и записи соответствующего у.о. узла СС в цепочку. С помощью специально разработанного языка сформулированы правила отображения, в соответствии с которыми функциональные слова отображаются в свои формализованные эквиваленты. Синтаксис языка имеет вид:

<правило>::=< антецедент> →< консеквент>

<антецедент>::= {<условие>}$^+$

<условие>::=(И{<выражение>}$^+$ /ИЛИ{<выраж.>}$^+$) /(НЕ<выраж.)

<консеквент>::=<тип взаимосвязи> /<узел СС> /<вид операции> /<позиция символа>

<тип взаимосвязи >::= @ /. /, /& / + /- / />*

<узел СС>::=<описание узла$_1$> /<опис. узла$_2$> /... /<опис. узла$_к$>

<вид операции>::=r /s /u /e

<позиция символа>::=1 /2 / ... /n, где в <> - имя синтаксической категории; знак ::= означает определение левой части выражения через правую; {}- определяет повторы; | – обозначает варианты; () – для группирования; $^+$ - означает присоединение, *<выражение>* – это логическое выражение со значениями истина или ложь представляет собой один из 4-х сформулированных в работе базисных типов утверждений; *<тип взаимосвязи>* – тип взаимосвязи, возвращаемый функцией I_f для записи в цепочку; *<узел СС>* - у. о. одного из $к$ понятий СС, возвращаемого функцией I_m для записи в цепочку; *<вид операции>* - одна из 4-х предложенных в работе операций по

нормализации текста на внешнем языке из множества {r, s, u, e}; *<позиция символа>* - позиция в цепочке, куда помещают анализируемый символ формализованного описания длиной *n*.

Разработан алгоритм автоматического кодирования ЕЯ информации, описанный на предлагаемом языке отображения.

В четвертой главе рассмотрены принципы и методы классификации формализованных описаний. Частные модели интерпретации разделены на однородные группы по числу групп описаний текста, например, модель для группы описания «Локализация»:

<Lok>=$\sum_{i=1}^{H}(\underset{j=1}{\overset{N}{M}} L_j)_i$; «Заболевания»: <Zab>=$\prod_{g=1}^{F} Z_g$ &<Lok>$_g$; «РТО»:

<Rto>=$\prod_{l=1}^{M} R_l \bullet \prod_{k=1}^{Q} S_k$ &<Lok>$_k \sum_{s=0}^{T} \pm \left(\prod_{v=1}^{W} R_v \right.$ &<Lok>$_v \left. \sum_{z=0}^{Y} * P_z \right.$ &<Lok>$_z)_{s+1}$,

где $\sum_{i=1}^{H}(\underset{j=1}{\overset{N}{M}} L_j)_i$ - цепочка из H подцепочек, соединенных взаимосвязью «перечисление»; $\underset{j=1}{\overset{N}{M}} L_j$ - цепочка длиной N, в которой соседние символы L соединены взаимосвязью «принадлежность»; $\prod_{k=1}^{Q} S_k$ - цепочка длиной Q, в которой соседние символы S соединены взаимосвязью «равноправие»; $\sum_{s=0}^{T} A_s$ - цепочка из T групп подцепочек A_s, соединенных взаимосвязью «перечисление», причем в случае $T = 0$ длина цепочки = 0; D_m^n - дерево с обозначением m и рангом n, который задает порядок следования медицинских понятий в формализованном описании,

$$R, P \in \{D_r, D_t\}, S \in D_1, L \in \{D_f^2, D_s^1, D_a^1, D_m^1, D_k^1, D_n^1, D_o^0, D_p^0, D_h^0\}, Z \in D_z$$

Примеры моделей интерпретации приведены с у. о. узлов СС в виде медицинских понятий, записанных в угловых скобках. Фрагменту текста **«Огнестрельное сквозное осколочное ранение правого плеча с переломом диафиза плечевой кости со смещением и повреждением лучевого нерва»** соответствует модель «РТО»: *<огнестрельный>•<сквозной>• <осколочный>• <ранение>& <правый>@<плечо>+<перелом>&<диафиз>@<плечевая кость>* ✶ *<смещение>&<лучевой нерв>* ✶ *<повреждение>&<лучевой нерв>*.

В медицинские тексты входят понятия, связанные с узлами СС разных уровней тематических деревьев. В работе выделены три случая соотношения уровней узлов в цепочках пациента и запроса и предложено использовать единый терминальный уровень анализа информации. Для этого все цепочки приводят к уровню описания, на котором символы цепочки - элементы основного словаря грамматики внутреннего языка, а каждому нетерминальному узлу поставлено в соответствие множество терминальных узлов. В случае пациента это множество может быть представлено в виде дизъюнкции* элементов (дизъюнктивное множество), в случае запроса - конъюнкции элементов (конъюнктивное множество): $P_i = \bigvee_{j=1}^{n} P'_{ij}$ - дизъюнктивное множество для i-го узла цепочки пациента с n терминальными потомками P'_j, и $Z_k = \bigwedge_{l=1}^{m} Z'_{kl}$ - конъюнктивное множество k-го узла цепочки запроса с m терминальными потомками.

В работе приведена и описана функциональная схема автоматизированной системы. После обработки ЕЯ сведений о пациенте, полученный набор цепочек проверяется на корректность и поступает в базу данных (БД), в которой накапливается информация по массиву пациентов. Аналогично запрос пользователя преобразуется в формализованное описание. При обработке система использует информацию из базы знаний о тезаурусе, грамматиках и т.д. Для получения ответов на запросы система производит вначале поиск по символам цепочки запроса, включенным в индексные файлы, с записью найденных сведений во временную БД, затем классификацию отобранных цепочек. В работе показано, что средствами одного поиска нельзя обеспечить необходимую точность выбора информации из БД, последовательный же анализ всех записей БД не удовлетворяет требованиям времени ответа на запрос, обосновано включение в индексные файлы только терминальных символов подгруппы "Локализация". На этапе классификации набор цепочек по каждому пациенту временной БД сопоставляют с цепочками запроса для принятия решения о соответствии описания пациента запросу.

Для выбора наиболее адекватного метода классификации в работе были исследованы подходы РО. Обоснована неприменимость стандартных методов для классификации цепочек. Сравнение цепочек

---

* Здесь и далее используются отдельные термины, заимствованные из алгебры логики, но отличающиеся тем, что они применены в области символьных операций.

пациента и запроса осложнено тем, что помимо синтаксического необходимо проверять и семантическое сходство, используя родовидовые отношения. Сформулированы следующие критерии релевантности описания пациента запросу:

1) цепочка пациента включает в себя все узлы или их потомки, которые указаны в цепочке запроса;
2) взаимосвязи между соответствующими семантическими группами медицинских понятий и внутри этих групп в цепочке запроса и цепочке пациента одинаковы.

В работе показано, что критерии смыслового соответствия сравниваемых цепочек тесно связаны с их синтаксисом. Эта связь положена в основу построенной алгебры цепочек, которая предназначена для определения релевантности описания пациента запросу. В качестве множества допустимых цепочек приняты цепочки, удовлетворяющие формуле (1), объединенные с множеством цепочек произвольного вида и цепочкой нулевой длины λ, принятой за единичную цепочку. Определены следующие операции алгебры:

1. **Дробление узла.**
2. **Расслоение составной цепочки.**
3. **Группировка.**
4. **Разбиение цепочки на подцепочки, соединенные жесткими взаимосвязями.**
5. **Посимвольное взятие дополнения цепочки.**
6. **Умножение простых цепочек.**
7. **Умножение составных цепочек.**
8. **Сравнение с символьным нулем.**

В работе подробно описаны операции, приведены примеры.

1. При приведении цепочки к терминальному уровню нетерминальные узлы представляют в виде множества терминальных (со штрихами). Цепочка пациента, состоящая из узлов P_1 и P_2, соединенных взаимосвязью C, представима как $P = P_1 C P_2$, где $P_1 = \{P'_{111}, P'_{112}, P'_{121}, P'_{1221}\}$, $P_2 = \{P'_{211}, P'_{212}, P'_{22}\}$. В результате дробления: $P = P'_{111}CP'_{211} \vee P'_{112}CP'_{211} \vee P'_{121}CP'_{211} \vee P'_{1221}CP'_{211} \ldots P'_{1221}CP'_{22}$, далее получают: $P = \bigvee_{\substack{i=1 \\ j=1}}^{\substack{m \\ n}} P'_i C P'_j$, где n, m - количество терминальных узлов у нетерминальных узлов P_i и P_j соответственно. В случае

запроса: $\bigwedge_{\substack{k=1 \\ l=1}}^{\substack{u \\ t}} Z_k^{'} C Z_l^{'}$, где t,u - количество терминальных узлов у нетерминальных узлов Z_k и Z_l соответственно.

2. При оперировании цепочками с более чем одной жесткой одноименной взаимосвязью (см. табл.) или с одной взаимосвязью «перечисление» (составные цепочки) возникают неоднозначные ситуации. Такие цепочки представимы в виде множества простых, в которых отсутствуют подобные взаимосвязи. Расслоение составной цепочки B_i - это преобразование цепочки в конъюнктивное множество n простых цепочек вида $B_{i1} \wedge B_{i2} \wedge ... \wedge B_{in}$. Например, составная цепочка *<осколочный>•<ранение>&<голова>,<область>@<ухо>* после расслоения будет преобразована в две простые: ***<осколочный>•<ранение>&<голова>*** **и** ***<осколочный>•<ранение>&<область>@<ухо>***

3. Третья операция основана на справедливости коммутативного закона. Пусть составная цепочка запроса после операций расслоения и дробления нетерминальных узлов представлена в виде комбинации простых цепочек: $(Z_{11} \wedge Z_{12}) \wedge (Z_{21} \wedge Z_{22} \wedge Z_{23})$. После группировки образуется конъюнктивное (для запроса) и дизъюнктивное (для пациента) множество, эквивалентное в информационном смысле исходной цепочке. Для данного случая:

$$(Z_{11} \wedge Z_{21}) \wedge (Z_{11} \wedge Z_{22}) \wedge (Z_{11} \wedge Z_{23}) \wedge (Z_{12} \wedge Z_{21}) \wedge (Z_{12} \wedge Z_{22}) \wedge (Z_{12} \wedge Z_{23})$$

Пример. Составная цепочка запроса имеет вид ***<ранение>& <верхняя конечность>,<нижняя конечность>.*** После расслоения получают две простые цепочки: 1) ***<ранение>& <верхняя конечность>,*** 2) ***<ранение>&<нижняя конечность>.*** После дробления нетерминальных узлов **<верхняя конечность>** и ***<нижняя конечность>*** получают для первой простой цепочки - *<ранение>&<плечо>* $\wedge$ *<ранение>&<предплечье>* $\wedge$ *<ранение>&<ладонь>* $\wedge$ *<ранение>&<запястье>* $\wedge$ *<ранение>&<кисть>,* для второй цепочки - *<ранение>&<бедро>* $\wedge$ *<ранение>&<голень>* $\wedge$ *<ранение>&<колено>* $\wedge$ *<ранение>&<подошва>* $\wedge$ *<ранение>&<пятка>* $\wedge$ *<ранение>&<стопа>.* Тогда конъюнктивное множество запроса: *{<ранение>&<плечо>* $\wedge$ *<ранение>&<бедро>; <ранение>&<плечо>* $\wedge$ *<ранение>&<голень> ... <ранение>&<предплечье>* $\wedge$ *<ранение>&<голень> ... <ранение>&<кисть>* $\wedge$ *<ранение>&<стопа>}*

4. Все взаимосвязи по их роли в цепочке разделены на жесткие и мягкие (см. табл.). Мягкие взаимосвязи допускают вариации местоположения узлов в цепочке без искажения ее смысла, жесткие – не допускают. После разбиения цепочка представлена как последовательность подстрок, соединенных жесткими взаимосвязями.

5. Пятая операция преобразует цепочку в строку, где каждый символ L представлен в виде своего дополнения $\overline{L}$. При перемножении таких символов получают единичную цепочку: $L \times \overline{L} = \lambda$.

6. Шестая операция умножения двух простых цепочек заключается в получении цепочки, у которой внутри каждой группы, разделенной жесткими взаимосвязями, перемножаются символы с соответствующими им дополнениями. Например, цепочки:

$Z_1 = \overline{R}_1 \overline{\vee} \overline{R}_2 \,\overline{\&}\, \overline{L}_6 \overline{\vee} \overline{R}_8 \,\overline{\&}\, \overline{L}_{12}$ и $Z_2 = \overline{R}_3 \overline{\vee} \overline{R}_1 \overline{\vee} \overline{R}_2 \,\overline{\&}\, \overline{L}_9$ умножают на $P = R_2 \cdot R_4 \cdot R_1 \,\&\, L_2 \,@\, L_6 \,@\, L_9 + R_7 \cdot R_8 \,\&\, L_4 \,@\, L_{12}$. В результате получают: $Z_1 \times P = R_4 \cdot L_2 \,@\, L_9 R_7 \cdot L_4 @$; $Z_2 \times P = \overline{R}_3 R_4 L_2 \,@\, L_6 \,@ + R_7 \cdot R_8 \,\&\, L_4 \,@\, L_{12}$.

7. Умножение составных цепочек. Пусть запрос - это составная цепочка, включающая две простые Z_1 и Z_2, цепочка пациента - это три простые цепочки P_1, P_2, P_3. Тогда произведение составных цепочек имеет вид: $(Z_1 \wedge Z_2) \times (P_1 \wedge P_2 \wedge P_3)$, т.е. сведено к умножению простых. Эмпирически получают следующее выражение:

$$(Z_1 \times P_1 \vee Z_1 \times P_2 \vee Z_1 \times P_3) \wedge (Z_2 \times P_1 \vee Z_2 \times P_2 \vee Z_2 \times P_3).$$

В общем виде: $Z \times P = \bigwedge_{i=1}^{n} \bigvee_{j=1}^{m} Z_i \times P_j$ (2), где n, m - число простых цепочек Z_i, P_j запроса и пациента соответственно.

8. Сравнение с единичной цепочкой (символьным нулем). Положительными цепочками считают такие, которые не содержат ни одного символа дополнения, отрицательными – цепочки, содержащие хотя бы одно дополнение.

Для разделения цепочек пациентов на два класса, релевантных и нерелевантных запросу, в работе введено понятие сходства. Сходство двух цепочек - это логическая переменная, принимающая значение "истина", когда сравниваемые цепочки принадлежат одному классу, "ложь" - в противном случае. Сходство цепочки пациента и цепочки запроса истинно, если истинно сходство хотя бы одного элемента из

дизъюнктивного множества терминальных цепочек пациента с одним из элементов конъюнктивного множества терминальных цепочек запроса. Сходство 2-х простых цепочек истинно, если в результате посимвольного взятия дополнения цепочки запроса и умножения на цепочку пациента получена неотрицательная цепочка. Сходство составных цепочек вычисляют на основе формулы (2).

Если Z, P - это множества соответственно цепочек запроса и пациента, то для определения релевантности пациента запросу определяют сходство каждой пары цепочек z_i, p_j декартова произведения $Z \times P$. Для этого паре z_i, p_j ставят в соответствие произведение $z_i \times p_j$ алгебры цепочек. Используют операцию умножения матриц. Множеству Z поставлена в соответствие матрица-столбец $\mathbf{Z}$, множеству P - матрица-строка $\mathbf{P}$. Произведение $\mathbf{Z} \times \mathbf{P} = \mathbf{R}$ - матрица размера $m \times n$. Каждый элемент матрицы - это произведение цепочек множеств Z, P, содержащих соответственно m, n терминальных цепочек.

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} \times \begin{bmatrix} p_1 & p_2 & \cdots & p_n \end{bmatrix} = \begin{bmatrix} z_1 p_1 & z_1 p_2 & \cdots & z_1 p_n \\ & \ddots & & \\ & & \ddots & \\ z_m p_1 & z_m p_2 & \cdots & z_m p_n \end{bmatrix}$$

Если хотя бы одна цепочка в паре z_i, p_j является составной, применяют 7-ю операцию. Сравнение каждого элемента матрицы с символьным нулем позволяет сделать заключение о сходстве цепочек пациента и запроса. Предложенный в работе алгоритм классификации создан на основе алгебры цепочек и позволяет определить принадлежность описания пациента запрашиваемому классу.

В работе приведены экспериментальные данные, обосновывающие преимущества построенной автоматизированной системы по сравнению с существующими системами путем оценки результатов ручной и автоматической обработки текстов, оценки качества классификации данных, подготовленных программой и врачом. Проанализированы причины ошибок при автоматическом способе кодирования и проведено сравнение с подобными системами. Сделан вывод, что результаты автоматической обработки не уступают зарубежным аналогам. Результаты классификации, полученные на обучающей (614 цепочек) и контрольной выборках (282 цепочки), показали: отклик системы равен 0,99, точность - 1,00.

ЗАКЛЮЧЕНИЕ

Основные результаты диссертационной работы:

1. Построен тезаурус предметной области, описывающий терминологию военной медицины.
2. Синтезирована лингвистическая модель медицинских сведений о пациенте на русском языке, позволяющая проводить формализацию информации в терминах тезауруса. Модель описана средствами индексной грамматики.
3. Построена модель интерпретации для описания формализованных медицинских сведений о пациенте. Модель имеет 5 разновидностей по числу лингвистических групп текста на профессиональном медицинском языке.
4. Предложен и реализован в автоматизированной системе метод лингвистической обработки медицинских документов, включающий три стадии: лексико-морфологический анализ, грамматический анализ и семантическую интерпретацию. На последней стадии осуществляется отображение естественно-языковой информации в формализованное представление.
5. Разработан язык отображения лингвистической модели на модель интерпретации, при помощи которого описана и программно реализована процедура кодирования. Результаты работы алгоритма кодирования на 896-ти медицинских документах пациентов показали высокое качество автоматического кодирования.
6. Предложена архитектура автоматизированной системы, осуществляющей обработку естественно-языковых сведений о пациентах с последующей формализацией и занесением в БД. По запросам пользователей система классифицирует информацию, хранящуюся в БД.
7. С целью проведения синтаксического анализа построена алгебра цепочек, позволяющая оперировать формализованными представлениями медицинских сведений и включающая 8 операций. В основу алгебры положена комбинация логических и структурных методов распознавания образов.
8. Разработан классификационный алгоритм для получения ответов на запросы пользователей путем выполнения алгебраических операций над формализованными описаниями. Апробация алгоритма на 896-ти реальных медицинских документах дала следующие результаты: отклик системы равен 0.99, точность поиска -1.00, что превышает показатели зарубежных аналогов.

Предложенные методы и алгоритмы разработаны с использованием материалов 10-ти тысяч историй болезни и реализованы в прототипе автоматизированной системы, эксплуатирующимся в Архиве военно-медицинских документов.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. **Крутов В.С., Трояновская О.В.** *АИПС в архиве Военно-медицинского музея*//Отеч. арх.- М.,1996.-№ 5.-С. 103-105.
2. **Крутов В.С., Трояновская О.В.** *Теоретические основы построения автоматизированной системы о раненых и больных*//Воен.-мед. журн. –1997. -№ 12. - С. 4-8.
3. **Крутов В.С., Трояновская О.В.** *Автоматизированная интерпретация персональной медицинской информации как альтернатива ручному кодированию*//Мед. информатика накануне 21 века: Тез. докл. Всерос. науч. конф., г. Санкт-Петербург, 27-29 мая 1997 г.- СПб, 1997.-С. 220-221.
4. **Трояновская О.В.** *Система искусственного интеллекта для автоматизированной обработки персональной медицинской информации*//Вопр. техн. обеспечения мед.-биол. исслед.: Сб. науч. тр..- СПб., 1998.- С. 43-48.- (Изв. ГЭТУ, вып. 518).
5. **Трояновская О.В.** *Лингвистические модели представления клинических данных на естественном языке для автоматизированного анализа медицинской информации* // Диагностика, информатика, метрология, экология, безопасность: Тез. докл. науч.-техн. конф., г.Санкт-Петербург, 30 июня–2 июля 1998.- СПб., 1998. - С.145-146.
6. **Трояновская О.В., Манило Л. А.** *Система перевода медицинских документов с естественного на формализованный язык*//Междунар. конф. по мягким вычислениям и измерениям-99:Сб. докл., г.Санкт-Петербург, 25-28 мая 1999. - СПб, 1999. Т.2.- С. 173-176.
7. **Трояновская О.В., Манило Л. А.** *Автоматизированный анализ медицинских документов*//Медико-эколог. информ. технологии: Тез. докл. 2-й Междунар. техн. конф., г. Курск, 19-21 мая 1999.-Курск, 1999.- С. 19 - 21.
8. **Troyanovskaya O., Manilo L.** *Personal Medical Information Classification in Compliance with Dynamic Formed Classes*//Biomedical Engineering & Medical Informatics'97: Proc. Intern. Workshop, Gliwice, Poland, September 2-5, 1997. - Gliwice, 1997. - P. 154-158.