Берсенева Е.А., Седов А.А.

ЭТАПЫ СОЗДАНИЯ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ЛЕКСИЧЕСКОГО АНАЛИЗА МЕДИЦИНСКИХ ДОКУМЕНТОВ (АИС «ЭЛЕКС»)

ФГБНУ «Национальный научно-исследовательский институт общественного здоровья им. Н.А. Семашко, Россия, Москва.

Berseneva E.A., Sedov A.A.

STAGES OF THE AUTOMATED SYSTEM LEXICAL ANALYSIS OF MEDICAL DOCUMENTS (AIS "ALEX")

National Research Institute for Public Health", Russia, Moscow

Резюме. В статье рассматриваются осуществленные этапы создания системы лексического анализа медицинских документов «Элекс» как средства повышения качества медицинских документов в условиях их электронного формирования. Также рассматриваются основные особенности созданной системы и перспективы ее дальнейшего развития.

Ключевые слова: качество медицинской документации, системы лексического анализа, семантический анализ, информационные технологии, комплексные автоматизированные информационные системы лечебно-профилактических учреждений, системы лексического контроля.

Abstract. This article discusses the steps implemented a system of lexical analysis of medical documents "Alex" as a means of improving the quality of medical records in terms of their electronic form. It also discusses the main features of the established system and the prospects for its further development.

Key words: the quality of medical records, lexical analysis system, semantic analysis, information technology, integrated automated information system of health care institutions, the lexical system of control.

На сегодняшний день компьютеры прочно вошли в ежедневную жизнь любого медицинского учреждения [1,2]

независимо от профиля оказываемых услуг, подчиненности и формы собственности, и невозможно представить формирование любого содержательного медицинского документа без использования компьютера.

Побочным эффектом данного процесса информатизации медицинских учреждений стало то, что теперь врач теперь имеет возможность взять готовый документ и просто «вставить» туда фамилию пациента, или «собрать» такой документ из фрагментов, подготовленных ранее, без связи с конкретным событием в клинической практике. При этом, в текущих условиях погоня за экономией времени в клинической практике подталкивает врача использовать наиболее простой путь для формирования содержательной части медицинского документа. Это приводит к тому, что снижается качество медицинских документов.

По нашему мнению, компьютеру должен быть противопоставлен компьютер. Осуществление контроля качества медицинской документации, просто вычитывая документы, крайне трудоемко и требует специального штата для решения этой задачи. Все вышеизложенное свидетельствует о безусловной практической необходимости в автоматизированной системе лексического контроля медицинских документов.

При реализации нами проекта по созданию автоматизированной системы лексического контроля был исследован новый подход к обработке русскоязычных медицинских текстов, включающий комплекс методов, как хорошо проверенных при решении сходных задач в других областях, так и перспективных подходов – поиск с учетом семантико-синтаксических структур предложений с учетом медицинской терминологии и специфичной лексики. Все разработанные нами и реализованные в автоматизированной системе «Элекс» подходы являются новыми и оригинальными в

части анализа медицинских текстов на русском языке.

На первом этапе проекта создания автоматизированной системы лексического анализа были получены следующие результаты. Созданы механизмы загрузки данных из текстового документа в виде, пригодном для пословного разбора, что позволяет осуществлять загрузку документов, поступающих из любых МИС. Механизмы представлены следующими сервисами: a. Сервис получения документов из файлового хранилища. Сервис осуществляет сбор документов по протоколу FTP, что позволяет взаимодействовать с информационными системами, не поддерживающими современные методы обмена данными. b. Сервис получения документов по SOA-модели взаимодействия. Предназначен для загрузки документов из систем, поддерживающих работу по с SOAP или REST сервисами. c. Сервис файлового хранения полученных документов в привязке к индексу. Получает и хранит первичные данные полученного документа в неизмененном виде.

Также были созданы механизмы пословного разбора текстовой информации, поступающей из электронной медицинской карты (ЭМК), что позволяет осуществлять первичную обработку документов, поступающих из любых МИС. Представлены следующими сервисами: a. Сервис создания индекса документов в базе данных. Регистрирует факт получения документа с учетом источника и стадии его обработки. В дальнейшем именно данные в индексе используются остальными сервисами системы в случае, если необходим доступ к документу. b. Сервис первичного разбора полученного документа. Производит разбор документа на отдельные элементы, с созданием соответствующих данных о документе в базе. Сервис выделяет фрагменты документа длинной в одно слово для сверки со словарем, длиной пять и десять слов – для сверки в сервисе структурного анализа.

Исследованы возможности применения систем анализа текстов на естественном языке (ЕЯ) для решения задачи разбора текстовой информации медицинского характера. Использование комплексного подхода, учитывающего лексико-морфологическую информацию слов текста, синтаксические связи, а также семантические значения слов и семантические связи между словами позволяет решать задачу автоматической оценки качества медицинского электронного документа (МЭД). В качестве системы морфологического и синтаксического анализа выбрана система АОТ [Автоматическая обработка текста. 2013. // [Электронный ресурс] — URL: http://www.aot.ru/], распространяемая под свободной лицензией LGPL. АОТ предоставляет функции нормализации слов ЕЯ и построения синтаксических структур предложения. Кроме того сформулированы принципы построения семантического анализатора, который должен решать задачу семантической разметки предложений ЕЯ-текстов. В его основу положена модель семантики ЕЯ Золотовой-Осипова – реляционно-ситуационная модель предложений текста [3]. Также предложен метод построения устойчивых словосочетаний на основе анализа частотности синтаксических конструкций в предложениях медицинских текстов. Получаемое в результате работы указанных методов представление текстовой информации является неоднородной семантической сетью (НСС).

Для сохранения результатов пословного разбора текстовой информации, поступающей из электронной медицинской карты (ЭМК), созданы механизмы и структуры данных в системе управления базами данных (СУБД). Они позволяют сохранять результаты обработки данных словарей медицинской и немедицинской лексики для первичного формирования соответствующих словарей. Реализация представляет собой базу данных содержимого документа,

словаря и типовых элементов. Все сервисы, преобразующие полученный документ, ведут запись в эту базу с привязкой к индексу документов. Построена на основе реляционной СУБД FireBird 2.5. База данных содержит следующую информацию: индекс документов, индекс слов словаря в документе, словарь, индекс чешуек в документе, чешуйки, справочник типов документов, справочник ftp-источников. Вне реляционной СУБД разработаны структуры данных, предназначенные для эффективного сопоставления информации текстов ЕЯ, включающие следующие индексы: индекс ключевой лексики документов (слов и устойчивых словосочетаний); индекс списковых структур для хранения объектов, их атрибутов и связей между объектами, в НСС; инвертированный поисковый индекс для реализации функций эффективного поиска в НСС, соответствующих текстам медицинской тематики.

На первом этапе проекта также был создан сервис структурного и лексического анализа. Он осуществляет сопоставление элементов, полученных в ходе работы предыдущего сервиса, с данными словаря и данными сервиса типовых элементов по «чешуйчатому» алгоритму. Практика показала наибольшую эффективность использования чешуек длиной пять и десять слов. Помимо этого модифицирован метод многокритериального сравнения текстов [4] с использованием лексико-морфологической информации слов текста, синтаксических связей, а также семантических значений слов и семантических связи между словами для решения задачи анализа первичных медицинских документов по структуре, содержанию, степени уникальности и количеству технических ошибок. Сопоставление текстов может производиться с учетом замены слов на синонимы и изменения порядка слов в предложении, не меняющего его смысла.

Создан сервис формирования реестра типовых элементов документа. Осуществляет формирование реестра

типовых элементов документа и сверку данных вновь поступающего документа с реестром с целью выявления так называемых «элементов шаблонов» - часто повторяющихся фрагментов документов, которые целесообразно отнести к структуре документа, а не к его содержательной части.

В рамках проекта создан интерфейс для первичной обработки данных в словарях медицинской и немедицинской лексики, который позволяет производить экспертную обработку содержимого соответствующих словарей. При создании интерфейса реализован сервис формирования словаря. Этот сервис используется для формирования словаря медицинской и немедицинской лексики. Словарь наполняется автоматически по результатам анализа поступающих документов, вручную производится только классификация элементов, что позволяет существенно снизить трудозатраты на его ведение. Каждое новое слово поступает оператору «на разбор». Оператор должен вручную отнести слово к одной из групп: а) медицинская лексика; б) общая лексика; в) «несловарное» слово. Оператор должен установить один из признаков: а) ссылка на родительскую словоформу (оставить пустым, если это слово и есть она); б) Орфографическая ошибка (если такая есть); в) область медицины. Также рассмотрены возможности применения методов автоматической классификации текстовой информации и разделения медицинской лексики на вышеописанные группы. Для этой цели исследованы практические возможности применения характеристики тематической значимости слов ЕЯ [5] к медицинским документам.

На втором этапе проекта получены следующие результаты. Построена модель алгоритмов вычисления критериев оценки количественного содержания медицинской информации в текстовом документе, поступающем в СУБД. Алгоритм вычисления критериев оценки количественного

содержания медицинской информации в текстовом документе основывается на выделении медицинских терминов в тексте и оценке их частотности. Алгоритм использует результаты синтаксического анализа текста медицинского документа и принимает во внимание обнаруженные в нем медицинские термины (в том числе составные). Оценка количественного содержания медицинской информации производится на основе анализа частотных характеристик всех слов и словосочетаний как в рассматриваемом тексте, так и в эталонной коллекции медицинских документов различных типов (анамнезы, эпикризы и т.п.). Алгоритм отделяет общеупотребительные медицинские термины от специфичных медицинских терминов. Доля информационной значимости (по TFIDF с нормировкой на 1) представляет собой оценку количественного содержания медицинской информации в текстовом документе.

Также сформулированы фундаментальные критерии оценки содержащейся в документе медицинской информации по степени уникальности. Для оценки степени уникальности (оригинальности формулировок) текста проверяемого медицинского документа предложено использовать пофрагментное сопоставление текста проверяемого с ранее созданными медицинскими текстами, имеющимися в индексной базе. Текстовая информация при этом представляется в виде неоднородной семантической сети (НСС). Предложены следующие фундаментальные критерии оценки содержащейся в документе медицинской информации по степени уникальности: А) Первый критерий заключается в оценке количественного «пословного» совпадения предложения проверяемого текста и предложений - потенциальных источников заимствований. Эта оценка представляет собой TFIDF-оценку с нормировкой на 1. Б) Второй критерий представляет собой оценку соответствия синтаксической структуры проверяемого предложения с предложениями-

источниками. В) Третий критерий основан на подсчете количества совпадающих семантических значений у соответственных слов (совпадающих по нормальной форме). Г) Четвёртый критерий базируется на подсчете количества совпадающих семантических связей между соответственными словами (совпадающих по нормальной форме) в проверяемом предложении и предложениях-источниках. В совокупности эти критерии позволяют обнаруживать дублирование медицинского текста, фрагменты которого могут являться как развернутыми распространенными предложениями, так и формальным описанием (например, перечислением симптомокомплексов).

Построена модель алгоритмов вычисления критериев оценки содержащейся в документе медицинской информации по степени уникальности. Алгоритмы вычисления критериев оценки содержащейся в документе медицинской информации по степени уникальности используют структуры данных, разработанные на первом этапе реализации настоящего проекта. Общий алгоритм оценки сходства предложений проверяемого медицинского документа включает следующие шаги: 1) Лингвистический анализ текста, построение НСС. 2) Выбор фрагментов-кандидатов на проверку степени уникальности. 3) Для каждого фрагмента: a) поиск, выборка и фильтрация информации из поисковых индексов; b) предварительная оценка сходства проверяемого фрагмента и предложений из индексной базы; c) оценка сходства проверяемого фрагмента и предложений из индексной базы на основе разработанных критериев А)-Г). 4) Оценка доли проверенных фрагментов, для которых превышено пороговое значение оценки сходства с фрагментами из индексной базы. Степень уникальности медицинского документа – есть доля заимствованных предложений в числе проверенных.

На последнем (третьем) этапе проекта была построена система автоматизированного семантического анализа

медицинских текстов любых видов, пригодная к применению как в клинической практике, так и в научной деятельности. Основными функциями системы являются следующие:

а) оценка количественного содержания медицинской информации в текстовом документе;

б) оценка содержащейся в документе медицинской информации по степени уникальности, которая опирается на фундаментальные критерии А)-Г), предложенные на втором этапе проекта;

в) оценка структуры документа, которая опирается на фундаментальные критерии оценки документа по структуре в соответствии с заявленным типом;

г) оценка содержащейся в документе медицинской информации по степени соответствия заявленному типу.

Разработанная система позволяет выполнять оценку медицинских документов различных типов (анамнезы, эпикризы и т.п.). В ходе опытной эксплуатации созданной системы автоматизированного семантического анализа медицинских текстов обработано более 70 000 документов, содержащих информацию медицинского характера.

Взаимодействие пользователей (заведующих отделениями, клинических экспертов, руководителей) с системой ведется через созданный интерфейс для экспертной и аналитической работы с использованием алгоритмов семантического анализа.

Литература

1. Берсенева Е.А. Информационные системы в управлении лечебно-профилактическим учреждением // Врач и информационные технологии. – 2006. - № 4. – С.75.

2. Берсенева Е.А., Седов А.А. Автоматизированный лексический контроль как средство повышения качества медицинских документов. // Менеджер здравоохранения. – 2014. - № 2. – С. 49-53.

3. Осипов Г.С. Приобретение знаний интеллектуальными системами. // М.: Наука. Физматлит, 1997.

4. Соченков И.В. Метод сравнения текстов для решения поисково-аналитических задач // Искусственный интеллект и принятие решений. М.: ИСА РАН. - №2. - 2013.- С.95-106.
5. Э. Мбайкоджи, А. А. Драль, И. В. Соченков. Метод автоматической классификации коротких текстовых сообщений. // Информационные технологии и вычислительные системы. – 2012. – №3. – с.93 – 102.

Вайсман Д.Ш.

АНАЛИЗ ВЛИЯНИЯ ОБУЧЕНИЯ ВРАЧЕЙ И ВНЕДРЕНИЯ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ НА ДОСТОВЕРНОСТЬ СТАТИСТИКИ СМЕРТНОСТИ

ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Минздрава РФ, Россия, Москва,

ФБГНУ «Национальный НИИ общественного здоровья имени Н.А. Семашко, Россия, Москва,

Vaisman D. Sh.

ANALYSIS OF THE IMPACT OF PHYSICIAN EDUCATION AND IMPLEMENTATION OF AUTOMATED CODING SYSTEM ON RELIABILITY OF MORTALITY STATISTICS

Federal Research Institute for Health Organization and Informatics of Ministry of Health of the Russian Federation, Moscow

National Research Institute for Public Health, Russia, Moscow

Резюме. Рассмотрены два фактора, влияющие на достоверность статистики смертности: обучение врачей и внедрение автоматизированной системы регистрации смертности. Проанализированы ошибки оформления свидетельств о смерти и