АВТОМАТИЗИРОВАННАЯ СИСТЕМА ЛЕКСИЧЕСКОГО АНАЛИЗА: ОСНОВНЫЕ МЕТОДЫ И ПОДХОДЫ

БЕРСЕНЕВА ЕВГЕНИЯ АЛЕКСАНДРОВНА, докт. мед. наук, руководитель Центра высшего и дополнительного профессионального образования ФГБНУ «Национальный научно-исследовательский институт общественного здоровья им. Н.А. Семашко», Россия, 105064, Москва, ул. Воронцово Поле, 12/1, стр. 1, тел. +7-916-216-84-59, e-mail: eaberseneva@gmail.com

СЕДОВ АЛЕКСАНДР АНАТОЛЬЕВИЧ, младший научный сотрудник Центра высшего и дополнительного профессионального образования ФГБНУ «Национальный научно-исследовательский институт общественного здоровья им. Н.А. Семашко», Россия, 105064, Москва, ул. Воронцово Поле, 12/1, стр. 1, тел. +7-903-624-64-51, e-mail: gd@seber.ru

Реферат. В статье рассматриваются актуальные вопросы создания автоматизированной системы лексического анализа медицинских документов. **Цель работы** — создание автоматизированной системы лексического анализа медицинских документов. **Материал и методы.** Адаптированные методы выявления нечетких дубликатов и заимствований в текстах, чешуйчатый алгоритм, реляционно-ситуационная модель, неоднородные семантические сети, многоэтапный лингвистический анализ. **Результаты и их обсуждение.** В рамках данного проекта 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, т.е. поиск семантико-синтаксических структур предложений с учетом медицинской терминологии и специфической лексики. **Заключение.** Разработанная система позволяет оценивать медицинские документы различных типов (протоколы осмотров, эпикризы и т.д.). Взаимодействие пользователей с системой ведется через созданный интерфейс для экспертной и аналитической работы с использованием алгоритмов семантического анализа. Применяемые алгоритмы опираются на созданные в рамках проекта словари медицинской и немедицинской лексики.

Ключевые слова: качество медицинской документации, системы лексического анализа, семантический анализ, информационные технологии, комплексные автоматизированные информационные системы медицинских организаций.

Для ссылки: Берсенева, Е.А. Автоматизированная система лексического анализа: основные методы и подходы / Е.А. Берсенева, А.А. Седов // Вестник современной клинической медицины. — 2017. — Т. 10, вып. 2. — С.7—12.

DOI: 10.20969/VSKM.2017.10(2).7-12.

LEXICAL ANALYSIS AUTOMATED SYSTEM: MAIN METHODS AND APPROACHES

BERSENEVA EVGENIA A., D. Med. Sci., Head of Higher and supplementary professional education department of N. Semashko National Research Institute of Public Health, Russia, 105064, Moscow, Vorontsovo Pole str., 12/1, bld. 1, tel. +7-916-216-84-59, e-mail: eaberseneva@gmail.com

SEDOV ALEXANDER A., junior research worker of Higher and supplementary professional education department of N. Semashko National Research Institute of Public Health, Russia, 105064, Moscow, Vorontsovo Pole str., 12/1, bld. 1. tel. +7-903-624-64-51, e-mail: gd@seber.ru

Abstract. Topical issues of the medical document lexical analysis automated system creation are considered in the article. **Aim.** Medical document lexical analysis automated system creation. **Material and methods.** Customized methods of indistinct counterpart and borrowings detection, scaly algorithm, relational and situation-dependent model, non-uniform semantic networks and multi-stage linguistic analysis. **Results and discussion.** The new approach to Russian medical texts processing that includes a complex of methods, both: well checked in similar problem solving in other areas as well as perspective approaches, such as search taking into account semantic syntactical sentence structures, medical terminology and specific vocabulary, was investigated in terms of the project. **Conclusions.** The system developed allows to perform an assessment of medical documents of different types (examination protocols, discharge summaries, etc.). Interaction between the users and system is carried via the interface for expert and analytical operation with the use of semantic analysis algorithms created. The algorithms applied are based on the medical and non medical vocabularies created within the project.

Key words: medical documentation quality, lexical analysis system, semantic analysis, information technologies, medical organization complex automated information systems.

For reference: Berseneva EA, Sedov AA. Lexical analysis automated system: main methods and approaches. The Bulletin of Contemporary Clinical Medicine. 2017; 10 (2): 7—12. **DOI:** 10.20969/VSKM.2017.10(2).7-12.

Информатизация здравоохранения в настоящее время является одним из приоритетных направлений развития отрасли. Текущую ситуацию с развитием информационных технологий в здравоохранении мы можем охарактеризовать следующим образом:

- Развитие информационных технологий в настоящее время находится на таком уровне, что позволяет решать самые сложные в технологическом плане задачи.
- К настоящему времени кардинально изменилась оснащенность медицинских организаций (МО) компьютерами.

Как видно из данных М.М. Эльянова (рис. 1) за период с 2000 по 2007 г. в структуре парка компьютеров МО превалировали компьютеры в бухгалтерии, отделе кадров, подразделении статистики и обязательного медицинского страхования (ОМС). Собственно в 2000 г. лечебно-диагностический процесс был слабо оснащен компьютерами, и ситуация не менялась в течение многих лет.

В настоящее время оснащение компьютерами превалирует в лечебно-диагностическом процессе, существенная доля персональных компьютеров (ПК) и вообще компьютерного оборудования (61,4%) используется для автоматизации именно лечебно-диагностического процесса [1].

Фактически, мы можем подвести итог тому большому периоду, когда компьютеры в основном применялись в МО для бухгалтерии, статистики и отдела кадров. Сейчас доля ПК хоть и высока (31,1%), но все же не главная. При этом лавинообразное внедрение информационных технологий ставит новые вопросы перед профессионалами в области медицинской информатизации, которых раньше просто не возникало.

Проведенное нами исследование показало, что качество существенной части документов, формируемых врачами с использованием средств вычислительной техники, неудовлетворительное. При этом следует заметить, что в большинстве случаев врач, работая на компьютере, не обладает навыком быстрой слепой печати. Объем документов при использовании компьютера существенно увеличился за счет использования так называемых «шаблонов».

Надо заметить, что под этим термином обычно понимается не электронный аналог бланка документа, а, скорее, полуфабрикат, уже содержащий фактические данные еще до того, как врач начал фиксировать информацию, относящуюся к пациенту, о котором повествует документ. Важным является именно это свойство «шаблона», которое совершенно извращает само понятие «шаблон», поскольку оно изначально предполагает отсутствие каких-либо связанных клинических данных.

Фактически же пациента еще не осмотрели, а данные в документе уже есть. Откуда они берутся? Обычно врачи формируют свою собственную «копилку» этих «шаблонов», складывая туда документы, содержащие наиболее характерные признаки, описание течения заболевания и т.д., а в момент формирования документа только вносят в шаблон небольшие правки (если не забывают) и сохраняют под видом вновь созданного документа. Тут врачи используют шаблоны, как им кажется, пациенту во благо. Особое свойство информации в электронном виде — это ее бесконечное размножение без разрушения первоисточника. Изучая эти особенности использования компьютеров в медицинской организации, мы обратили более пристальное внимание на содержание документов, формируемых на компьютере. Более подробное знакомство с результатами использования таких «шаблонов» однозначно и неутешительно.

Качество существенной части документов неудовлетворительное. Причем не только в части отображения деталей течения, микросимптомов, редких диагностических находок, промежуточных выводов, задач для дальнейшего наблюдения, но и в части совершенно формальной. Применение «шаблонов» приводит не только и не столько к нарушениям отдельных элементов документа, сколько к разрушению именно смысловой целостности документа и утрате описания лечебно-диагностического процесса как такового, что в дальнейшем не дает возможности использовать документ как самостоятельную информационную сущность для рассмотрения. Другими словами, компьютер используется только для ввода информации, но никак не используется для вывода (анализа).

Рис. 1. Структура парка компьютеров в медицинских организациях (данные М.М. Эльянова)

Хотя неоднократно показано, что именно в части вывода и анализа накопленной информации компьютеры оказываются наиболее полезными. Иначе они выступают только в роли «высокотехнологичной печатной машинки с доступом в Интернет».

На наш взгляд, единственно правильным решением в данной ситуации будет противопоставить компьютеру компьютер. Другой компьютер, который выступит на стороне главного врача, заведующего отделением, клиникой, а зачастую и здравого смысла, эффективно проведет формальную оценку подготовленного документа.

Таким образом, появление автоматизированной системы лексического анализа и ее использование в медицинских организациях, безусловно, насущно и актуально.

Вопросы анализа медицинских данных, представленных в том числе в текстовой форме, относятся к активно исследуемой и развиваемой области научных знаний. Однако в России в настоящее время отсутствуют работы в области адаптации лингвистических методов анализа текстов, оценки качества медицинских текстов, в том числе поиска фрагментных заимствований, оценки оригинальности текстового содержания этих документов, оценки структуры и соответствия содержащейся в документе информации типу медицинского документа.

В рамках нашей работы 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, т.е. поиск семантико-синтаксических структур предложений с учетом медицинской терминологии и специфической лексики.

На первом этапе проекта было реализовано следующее. Созданы механизмы загрузки данных из текстового документа в виде, пригодном для пословного разбора, что позволяет осуществлять загрузку документов, поступающих из любых медицинских информационных систем (МИС). Данные механизмы в настоящий момент представлены следующими сервисами:

а. 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.

б. Сервис получения документов по SOA-модели взаимодействия. Предназначен для загрузки документов из систем, поддерживающих работу с SOAP- или REST-сервисами.

с. Сервис файлового хранения полученных документов в привязке к индексу. Получает и хранит первичные данные полученного документа в неизменном виде.

Также созданы механизмы пословного разбора текстовой информации, поступающей из электронной медицинской карты (ЭМК), что позволяет осуществлять первичную обработку документов,

поступающих из любых МИС. Данные механизмы представлены следующими сервисами:

а. Сервис создания индекса документов в базе данных, который 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.

б. Сервис первичного разбора полученного документа, который производит разбор документа на отдельные элементы, с созданием соответствующих данных о документе в базе. Сервис выделяет фрагменты документа длиной в одно слово для сверки со словарем и длиной в пять и десять слов — для сверки в сервисе структурного анализа.

Также были 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 AOT (<http://www/aot.ru>), распространяемая под свободной лицензией LGPL. AOT 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 естественного языка Золотовой — Осипова — реляционно-ситуационная модель предложений текста [2].

Кроме того, 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 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. Построенная 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, предназначенных для эффективного сопо-

ставления 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, соответствующий текстам медицинской тематики.

На первом этапе проекта также создан сервис структурного и лексического анализа. Данный сервис 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 в пять и десять слов. Помимо этого, модифицирован метод многокритериального сравнения текстов [3] с использованием лексико-морфологической информации слов текста, синтаксических связей, а также семантических значений слов и семантических связей 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.

Кроме того, создан сервис формирования реестра типовых элементов документа. Данный сервис 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 этой работы создан интерфейс для первичной обработки медицинской и немедицинской лексики в словарях, интерфейс 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 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 [4] 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 специфических медицинских терминов. Доля информационной значимости [по алгоритму Text Frequency Inverse Documentary Frequency (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 проверяемого текста 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 соответствующих слов (совпадающих по нормальной форме).

Г. Четвертый критерий базируется на подсчете количества совпадающих семантических связей между соответствующими 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: а) поиск, выборка и фильтрация информации из поисковых индексов; б) предварительная оценка сходства проверяемого фрагмента и предложений из индексной базы; с) 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 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 (рис. 2, 3).

Рис. 2. Основное диалоговое окно автоматизированной информационной системы (АИС) «Элекс», представляющее собой настройку основных алгоритмов системы

Рис. 3. Диалоговое окно АИС «Элекс», предоставляющее возможности настройки поиска произвольного текста

В настоящее время проводятся работы по дальнейшему расширению практического применения АИС «Элекс» в медицинских организациях.

Прозрачность исследования. Исследование не имело спонсорской поддержки. Авторы несут полную ответственность за предоставление окончательной версии рукописи в печать.

Декларация о финансовых и других взаимоотношениях. Все авторы принимали участие в разработке концепции, дизайна исследования и в написании рукописи. Окончательная версия рукописи была одобрена всеми авторами. Авторы не получали гонорар за исследование.

ЛИТЕРАТУРА

1. Оценка оснащённости компьютерным оборудованием медицинских организаций в Российской Федерации / Е.П. Какорина, А.В. Поликарпов, Е.В. Огрызко, Т.Ю. Голубева // Менеджер здравоохранения. — 2015. — № 8. — С. 49—56.
2. Осипов, Г.С. Приобретение знаний интеллектуальными системами / Г.С. Осипов. — М.: Наука; Физматлит, 1997. — 109 с.
3. Соченков, И.В. Метод сравнения текстов для решения поисково-аналитических задач / И.В. Соченков // Искусственный интеллект и принятие решений. — 2013. — № 2. — С. 95—106.

4. Мбайкоджи, Э. Метод автоматической классификации коротких текстовых сообщений / Э. Мбайкоджи, А.А. Драль, И.В. Соченков // Информационные технологии и вычислительные системы. — 2012. — № 3. — С. 93—102.

REFERENCES

1. Kakorina EP, Polikarpov AV, Ogryzko EV, Golubeva TJu. Ocenka osnashhennosti komp'yuternym oborudovaniem medicinskih organizacij v Rossijskoj Federacii [Computer equipment assessment of the medical organizations in the Russian Federation]. Menedzher zdravoohraneniya [Healthcare Manager]. 2015; 8: 49-56.
2. Osipov GS. Priobretenie znanij intellektual'nymi sistemami [Acquisition of knowledge intellectual systems]. Moskva: Nauka; Fizmatlit [Moscow: the science; Fizmatlit]. 1997: 109 p.
3. Sochenkov IV. Metod sravneniya tekstov dlja resheniya poiskovo-analiticheskikh zadach [Method of texts comparison for the solution of search and analytical tasks]. Iskustvennyj intellekt i prinjatje reshenij [Artificial intelligence and decision-making]. 2013; 2; 95-106.
4. Mbajkodzhi Je, Dral' AA, Sochenkov IV. Metod avtomaticheskoy klassifikacii korotkih tekstovyh soobshhenij [Method of automatic classification of short text messages]. Informacionnye tehnologii i vychislitel'nye sistemy [Information technologies and computing systems]. 2012; 3: 93-102.