Системный анализ в медицине

Лингвистический анализ историй болезни для выявления факторов риска инсульта*

Н.А. Благосклонов, В.В. Донитова, Д.А. Киреев, Б.А. Кобринский, И.В. Смирнов

Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» РАН», г. Москва, Россия

Аннотация. Выявление и оценка факторов риска заболеваний необходимы для повышения эффективности профилактических мероприятий. Большое значение это имеет в отношении такой социально значимой патологии как инсульт. Применение автоматизированных методов для анализа больших массивов историй болезни может повысить эффективность извлечения информации о факторах риска, что показано в данном исследовании с использованием разработанных правил и лингвистического парсера.

Ключевые слова: *лингвистический парсер, разметка текста, факторы риска, инсульт.*
DOI: 10.14357/20790279200309

Введение

Известный в настоящее время перечень факторов риска инсульта опирается на различные исследования, проведенные в разных странах [1–3] Среди них как результаты длительного наблюдения за отдельными категориями больных, так и выборочные исследования больных с хронической ишемией мозга и инсультами. Можно отметить значительные отличия в частоте отдельных факторов риска, что определяется и различиями анализируемых этнических и возрастных групп.

Среди многочисленных факторов риска (предикторов заболевания), представленных в публикациях, можно выделить наиболее важные. В то же время, практически отсутствуют большие данные, опирающиеся на многие тысячи случаев наступившего инсульта, основанные на анализе персональных (индивидуальных) историй болезни российских специализированных неврологических отделений клинических больниц. Однако наиболее важные факторы риска, привлекающие внимание пациентов и врачей, следует искать в анамнезе и клинической картине пациентов с хронической ишемией мозга, проходящих лечение в стационарах, и госпитализируемых с инсультом. Из этого проистекает необходимость целевого анализа данных в массивах текстов. В связи с этим для настоящего исследования были составлены перечни потенциальных предикторов. Эти факторы (признаки) требуют обнаружения в тексте историй болезни для последующего анализа. Такое исследование может быть основано на предварительном их выявлении в ручном режиме во врачебных записях и последующем автоматическом поиске в историях болезней с помощью специальных методов интеллектуального анализа текстов. Это даст возможность в дальнейшем, используя средства интеллектуального анализа данных и, в частности, машинного обучения [4,5], решать задачи прогнозирования течения ишемической болезни мозга и опасность острого нарушения мозгового кровообращения.

* Работа выполнена при поддержке РФФИ в рамках научного проекта № 19-29-01090 мк.

Извлечение информации из медицинских текстов может быть автоматизировано несколькими методами. Одним из таких методов, рассматриваемых в данной статье, является создание специализированных правил-шаблонов для каждого термина на основе заранее размеченных текстов. Далее, на основе правил, создается лингвистический парсер, способный по ключевым словам и фразам выявить тексты, в которых употреблялся данный термин. Такой подход позволяет уменьшить количество необходимых для разметки текстов, так как он напрямую зависит от предварительного анализа экспертом и не требует обучения. Но у такого подхода есть несколько недостатков, например, переобучение, когда созданные правила слишком сконцентрированы на рассматриваемом корпусе, и очень узкая спецификация: если какой-то вариант написания не встречался в рассматриваемом датасете (наборе данных), то для него не будет правил и, следовательно, он не будет обнаружен. Для решения данных проблем можно воспользоваться методами активного обучения, при которых не требуется размечать все тексты и достаточно начать с небольшого количества размеченных текстов, и размечать только необходимые термины во время обучения модели, как было описано в [6]. Такой подход позволяет сократить количество необходимых для разметки текстов и сущностей. Важно отметить, что оба метода специализированы на анализе текстов, где для разметки необходимо иметь глубокие знания о размечиваемых терминах. Если рассматриваются термины, не требующие специализированных знаний, возможно использование методов на основе нейронных сетей для извлечения информации с использованием большого корпуса размеченных текстов. Такой корпус возможно создать самостоятельно, так как он не требует экспертов, или воспользоваться одним из доступных для исследователя. На уже имеющемся корпусе можно дообучить модель по извлечению именованных сущностей, как описано в [7].

Кроме того, стоит отметить, что при разметке текстов возможно указание связи между терминами, что увеличивает точность оценивания [8].

1. Методы и материал исследования

На первом этапе работы для создания текстовых аннотаций, то есть для добавления заметок к существующим текстовым документам, использовался веб-инструмент BRAT (http://brat.nlplab.org/index.html). BRAT предназначен, в частности, для структурированных аннотаций, где заметки не являются произвольным текстом, а имеют фиксированную форму (https://brat.nlplab.org/configuration.html). Данная структура может автоматически обрабатываться и интерпретироваться. Иерархия интерпретируется системой как таксономия.

Структура аннотации BRAT контролируется текстовыми файлами конфигурации, которые можно создавать и редактировать в любом текстовом редакторе. Файлы конфигурации имеют простую линейно-ориентированную структуру и синтаксис, знакомый по многим другим текстовым системам конфигурации. Организация верхнего уровня файлов конфигурации состоит из разделов, каждый из которых помечается строкой, содержащей только «[имя-раздела]», которое является одним из набора предварительно определенных имен разделов, определенных для файла конфигурации. Внутри разделов каждая непустая строка определяет один элемент конфигурации, где с первой непустой последовательностью именуется элемент, а остальная часть строки-спецификация, которая зависит от раздела. Пустые строки игнорируются.

Каждая строка в разделе [relations] определяет тип отношения, который может связать сущности, и, необязательно, свойства отношения. Отношения часто могут принимать аргументы более чем одного типа.

Раздел [attribute] определяет двоичные или многозначные «флаги», которые можно использовать для маркировки других аннотаций. Например, атрибут Negated может использоваться для пометки вхождения аннотированного события как явно запрещенного в тексте, или атрибут Confidence для пометки события как определенного, вероятного или сомнительного. Каждая строка в разделе [attribute] определяет тип атрибута и аннотации, к которым он может прикрепляться. Для двоичных атрибутов возможные значения true и false являются неявными. Для многозначных атрибутов должны быть указаны возможные значения. Тип (или типы) аннотаций, к которым может применяться атрибут, определяются с использованием синтаксиса ARG: TYPE, который также используется в определении отношений и событий (ARG, как правило, «Arg»). Значения, которые может принимать многозначный атрибут, определяются с использованием синтаксиса Value: VAL1 | VAL2 | VAL3 [...], где «Value» – это буквенная строка, а VAL1, VAL2 и т.д. – возможные значения.

На втором этапе был применен специально разработанный программный продукт, основанный на правилах для автоматического анализа медицинских текстов. Каждое правило представляет собой множество вариантов написания термина, соответствующего некоторой сущности. Правила

описывают как контекст используемого термина, так и ключевые слова, но могут включать и отрицательные признаки. Например, для термина «ожирение» положительными признаками являются слова «ожирение» или «повышенного питания», а отрицательными: «нормального питания».

Для анализа текстов с помощью правил использовалась библиотека для python yargy (https://github.com/natasha/yargy). Единственным минусом данной библиотеки является низкая скорость, но данный недостаток компенсируется простотой использования и большим количеством возможностей. По ранее составленным правилам были созданы специальные конструкции, которые могут определять, встречается ли конкретный термин в тексте и его значение. Самой часто используемой конструкцией была та, которая создает конвейер (pipeline) из нормализованной формы переданных слов, включенных в правила. Также использовались регулярные выражения с их реализацией из стандартной библиотеки python.

Материалом исследования являлись 6907 деперсонифицированных (обезличенных) неструктурированных электронных историй болезней пациентов неврологического отделения Федерального научного клинического центра Федерального медико-биологического агентства. Одним из критериев отбора историй болезни было обязательное наличие у пациента истории госпитализаций, т.е. количества госпитализаций больше единицы. Сбор информации о госпитализациях пациентов, предшествующих проявлению нарушений мозгового кровообращения, обусловлен необходимостью получения данных о ранее существовавших факторах риска. Таким образом, имея срез по каждому пациенту, мы получили возможность оценить, как менялись наличие и выраженность факторов риска в различные возрастные периоды жизни больного. Это позволило изучить влияние состояния здоровья пациентов (в анамнезе) и образа их жизни на развитие инсульта.

2. Выбор факторов риска для исследования

Сначала был определен перечень факторов риска, характеризующихся высокой угрозой развития инсульта. Они были выделены среди многочисленных встречающихся в литературе факторов риска, или предикторов инсульта [9,10]. С учетом важности и доступности получения этой информации из историй болезни на данном этапе исследования был отобран ряд факторов риска: гипертензия (включая повышение артериального давления при отсутствии установленного диагноза гипертонической болезни), головокружение, головная боль, мигрень, атеросклероз, ожирение. Обоснованием этого выбора стали исследования, проведенные в разных странах.

Результаты проведенного в Японии исследования показали, что артериальная гипертензия постоянно ассоциируется с повышенным риском инсульта, в то время как вклад остальных «традиционных» факторов риска авторами на их материале не подтвердился [11]. Проведенные мета-анализы также показали ассоциацию инсульта с гипертонией [12] и с индексом массы тела [13]. Для вертебробазилярной недостаточности, характеризующей ишемию мозга, характерны частые приступы головокружений системного и несистемного характера, а также головные боли, распространяющиеся от затылка ко лбу («жест снимания каски») [14,15]. Факторы риска раннего и прогрессирующего атеросклероза изучались в процессе эпидемиологического исследования и оценивались с использованием анализа логистической регрессии [16]. С поправкой на возраст, пол и этническую принадлежность, ожирение (ИМТ> 30 кг/м^2) связано с повышенным риском инсульта (отношение шансов 1,57, 95% ДИ = 1,28–1,94) [17].

3. Первичная обработка реальных данных

Для выделения необходимых данных из текста в качестве обучающей выборки использовались данные 106 историй болезни, размеченных врачами-экспертами и когнитологами. Предварительно был подготовлен перечень терминов, включающий все возможные варианты написания факторов в историях болезни и их интервальные значения. В медицинской документации отсутствует единый подход к формулировкам терминов. В связи с этим перечисленные выше факторы риска характеризуются многочисленными вариантами, включая косвенную информацию о приеме лекарств, указывающих на наличие определенного заболевания.

Соответственно, в историях болезни при ручной разметке были выявлены варианты терминов (в том числе сокращения слов), относящихся к факторам риска. Ниже приведены варианты для некоторых из них.

Атеросклероз встречался как: атеросклероз, атеросклеротический.

Гипертензия и повышение артериального давления встречались как: гипертензия, гипертоническая болезнь, артериальная гипертензия, артериальная гипертония, АГ, ГБ, гипертензивная болезнь, гипотензивная терапия, антигипертензивная терапия, принимает гипотензивные препараты, прини-

мает антигипертензивные препараты, АД рабочее, адаптирован к ### мм. рт. ст., привычные цифры АД ###/###, АД максимальное, мах АД, макс. АД, макс ### мм рт. ст., в настоящее время АД повышается до ### мм рт. ст., периодически повышение артериального давления, максимально до ### мм рт. ст., повышение цифр АД до ### мм рт. ст. максимально, с максимальными цифрами АД до ### мм рт. ст., повышение цифр АД до ### мм рт. ст., повышение АД до ### мм рт. ст., подъемы АД, АД ### мм рт. ст., артериальное давление, САД, ДАД, АД. В то же время необходимо отметить, что ряд сходных сокращений инструментальных и лабораторных исследований не указывает на наличие гипертензии: МАГ, ДСМАГ, КАГ, ФАГ, ААГ, выполнение АГ (сокращенное обозначение метода артериографии), АГ брюшной аорты и различные аббревиатуры в разделе «Лабораторные исследования».

Головокружение встречалось как: головокружение, вестибуло-мозжечковый синдром, вестибуломозжечковый синдром, вестибуло-атактический синдром, вестибулоатактический синдром, вестибулоатаксический синдром, вестибуло-атаксический синдром, вестибуло-кохлеарный синдром, вестибулокохлеарный синдром, вестибулопатия.

Головная боль встречалась как: головная боль, боль в затылочной области, боль в теменной области, боль в заушной области, боль в височной области, краниалгия, цефалгия, цефалгический синдром.

Мигрень встречалась как: мигрень, мигренозный.

Ожирение встречалось как: ожирение, алиментарное, алиментарно-конституциональное, экзогенно-конституциональное, избыточное развитие подкожно-жировой клетчатки, избыточное развитие подкожно-жирового слоя, прибавка массы тела, распределение подкожно-жировой клетчатки по абдоминальному типу, распределение подкожно-жировой клетчатки по смешанному типу, повышенного питания, избыточного питания. Отсутствие ожирения фиксировалось как: нормального питания, удовлетворительного питания, умеренного питания, пониженного питания.

Также необходимо отметить, что слово «синдром» могло быть записано в сокращенной форме: с-м. Кроме того, в историях болезни в перечисленных терминах встречались опечатки, перестановки и пропуски знаков, и другие искажения.

4. Обработка текстов историй болезни

После выделения терминов, рассматриваемых как потенциальные факторы риска инсульта, была разработана «структура разметки», которая определяла сущности (т.е. факторы риска), их атрибуты, связи между сущностями. Конкретная структура разметки создавалась внутри конфигурационного файла разметчика, который составляется заново для каждой новой задачи. Файл включает обязательные разделы, которые должны быть указаны. В том случае, если какой-то из разделов [entities], [relations], [attributes], [events] не нужен, он все равно должен присутствовать, но остается пустым. В данной работе не использовался раздел [events] – события.

Рассмотрим каждый из разделов созданного конфигурационного файла.

Раздел сущности [entities]. Условно в данной работе он представлен в виде многоуровневого списка, так как иерархическая вложенная структура необходима для классификации выделяемых сущностей и для удобства работы. В случае использования одноуровневого списка, его постоянно надо было бы листать, так как он был бы очень длинным. Иерархия позволяет сначала отображать верхние уровни и затем, раскрывая их, достигнуть интересующей нас сущности. Однако следует отметить, что термины в текстах размечаются не только с помощью самых нижних уровней иерархии, но также имеется возможность разметить термин, используя название раздела. Из примера, приведенного ниже, это соответствует ситуации, когда в тексте встречается название заболевания, которое не было введено в структуру разметки, но его необходимо выделить, чтобы потом создать под него сущность. Для этого можно воспользоваться разметкой термина с помощью названия раздела «Заболевания в анамнезе». Таким образом концептуально каждая строка в данном списке является сущностью.

В общем виде пример структуры сущностей выглядит следующим образом:

- Название раздела
  - Сущность 1
    - Вариант классификации сущности
      - Значение 1
      - …
      - Значение n
    - Вариант классификации сущности 2
  - Сущность 2

Примером конкретного файла со структурой может служить следующий фрагмент (варианты на английском и русском языках):

- Medical_history
  - Arterial_hypertension
    - Arterial_hypertension_stage

    - Arterial_hypertension_degree
    - Arterial_hypertension_risk
  - Atherosclerosis
- Заболевания в анамнезе
  - Артериальная гипертензия
    - Стадия
    - Степень
    - Риск
  - Атеросклероз

Раздел атрибуты [attributes]. В случаях, где это необходимо, была добавлена возможность выделения числовых и лингвистических характеристик сущностей с помощью опции атрибут. Для этого после списка сущностей необходимо было указать атрибуты, характерные для определенных терминов.

В общем виде это выглядит следующим образом:

Имя_атрибута Arg: *Сущность_1|Сущность_2|Сущность_n*, Value:*Value_1|Value_2|Value_n*

Имя атрибута обязательно должно сопровождаться его названием. *Сущность_1*, …, *Сущность_n* должны включать перечисления названий сущностей из предыдущего раздела ([entities]), у которых мог быть атрибут с данным названием. *Value_1*, …, *Value_n, что* предполагает указание на заранее предопределенные возможные значения данного атрибута; если заранее не определено какими могут быть значения, то указывается просто Value:Value.

На примере артериальной гипертензии это выглядит следующим образом:

Hypertension Arg:Arterial_hypertension|Arterial_hypertension_degree|Arterial_hypertension_stage|Arterial_hypertension_risk, Value:Value

Это означает: атрибут Hypertension может быть у сущностей артериальная гипертензия, степень артериальной гипертензии, стадия артериальной гипертензии и риск артериальной гипертензии. Данный атрибут может быть заранее не предопределен, в связи с тем, что с его помощью предполагается размечать числовые (в некоторых случаях лингвистические) характеристики, в том числе арабские или римские (например, iii или Ш вместо 3).

Между сущностями и атрибутами дополнительно связи прописывать не требуется, так как разметчик «понимает», что они уже как-то связаны.

Раздел связи [relations]. В данном разделе указываются типы связей, необходимых для осуществления разметки, и сущности, между которыми данная связь может быть осуществлена.

В общем виде задание связей выглядит следующим образом:

Имя_связи: Arg1:
Сущность_1. Arg2: C*ущность_2.*

Имя_связи обязательно необходимо задать. Arg1 – сущность-родитель (сущность, от которой идет связь); Arg2 – сущность-ребенок (сущность, к которой идет связь). Направление связи может быть задано и в другом направлении.

В настоящей работе введена связь с названием *Имеет_атрибут* (*Has_attribute_rel*).

Реальный пример: Has_attribute_rel Arg1:Arterial_hypertension, Arg2:Arterial_hypertension_stage|Arterial_hypertension_degree|Arterial_hypertension_risk.

То есть сущность «артериальная гипертензия» имеет следующие атрибуты (в виде самостоятельных сущностей): степень, стадия, риск. Пример разметки представлен на рис. 1.

На этом закончился этап работы по созданию системного файла со структурой разметки. Следующий этап работы заключался в обработке массива историй болезни. В табличном файле каждая строка – это история болезни, соответствующая отдельной госпитализации, а колонки – отобранные заранее поля из документов. Для разметчика BRAT необходимо было провести предварительную обработку данного массива информации, а именно: каждую строку представить в виде отдельного текстового файла с расширением *.txt. Это было осуществлено с помощью специально разработанного скрипта на языке программирования Python. Каждому файлу было присвоено имя следующим образом:

patient_number_XXX.0_number_YYY.0,

где XXX – ID пациента, а YYY – ID истории болезни.

В связи с требованиями разметчика BRAT были дополнительно созданы пустые файлы с расширением *.ann для каждой истории болезни. Это необходимо для того, чтобы в последующем в данные файлы записывалась вся размеченная информация.

Следующий этап работы был связан с разметкой и выгрузкой. В процессе разметки было необходимо в разметчике открыть интересующий текстовый файл и после нахождения соответствующего термина выделить необходимое слово или словосочетание. Затем в открывшемся контекстном меню осуществлялся выбор сущности, и если размечаемое слово являлось атрибутом или значением атрибута, то это указывалось. При выделении нескольких сущностей в тексте можно было установить между ними отношения.

После того, как истории болезни были размечены экспертами, была произведена выгрузка текстовых файлов и файлов аннотаций, содержащих

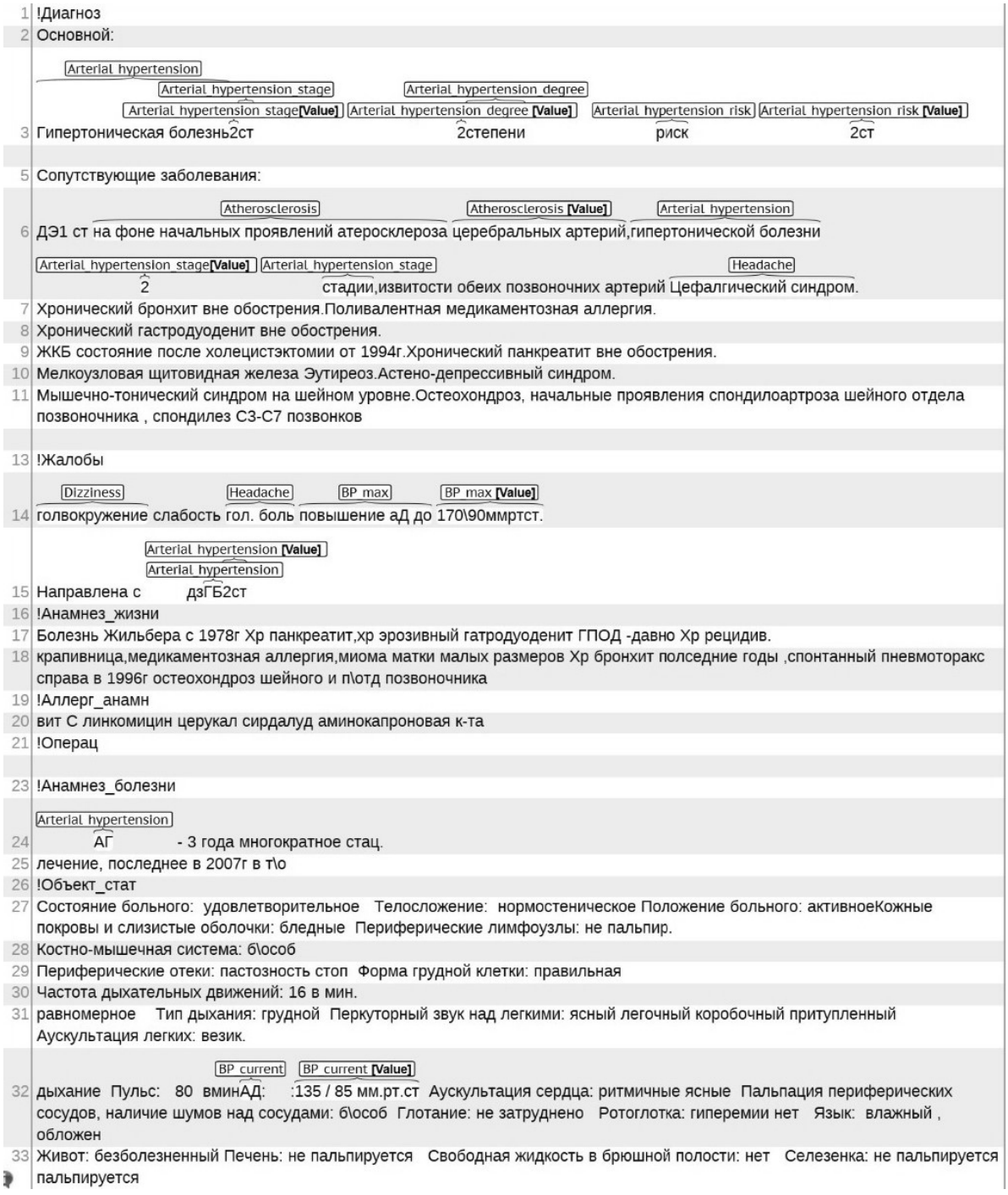

Рис. 1. Пример разметки медицинского текста в разметчике BRAT

разметку. Пример фрагмента разметки (в отношении артериальной гипертензии) из файла *.ann выглядит следующим образом:

T1 ...Arterial_hypertension 26 49 ..... Гипертоническая болезнь
T4 ...Arterial_hypertension_risk 62 66 ..... риск
T5 ...Arterial_hypertension_risk 67 70 ..... 2ст
A3...Hypertension T5
T2 ...Arterial_hypertension_stage 50 52 .... ст
T3 ...Arterial_hypertension_stage 49 50 2
A1...Hypertension T3
T6 ...Arterial_hypertension_degree 54 61 .. степени
T7 ...Arterial_hypertension_degree 53 54 .. 2
A2...Hypertension T7

В файле с разметкой используются следующие обозначения для строк:

- Если строка начинается с T:
  - T# – сущность с номером;
  - Название сущности;
  - Номер символа перед первым символом термина;
  - Номер последнего символа термина;
  - Выделенный термин;
- Если строка начинается с A:
  - A# – атрибут с номером
  - Название атрибута
  - Номер сущности, к которой относится атрибут.

То есть в приведенном примере в первой строке записано следующее:
T1 – номер сущности; Arterial_hypertension – название сущности; 26 – номер символа перед буквой «Г» в слове «Гипертоническая»; 49 – номер последнего символа в слове «болезнь», то есть буква «ь»; «Гипертоническая болезнь» – размеченный термин.

На примере седьмой строки:
A1 – номер атрибута; Hypertension – название атрибута; T3 – номер сущности, которая является атрибутом. В данном примере 2 – номер стадии артериальной гипертензии, которая была размечена как атрибут со значением.

Следует отметить, что система BRAT размечает что-то как сущность, а потом ставит пометку, что это на самом деле атрибут (значение атрибута).

Таким образом отражены сущности (с их названиями) и атрибуты.

5. Результаты проверки правил

Востребованные задачи автоматической обработки текстов на естественных языках включают в себя поиск по запросу, классификацию и кластеризацию текстов, извлечение знаний и фактов из текстов, поиск близких текстов и многие другие задачи. Качество результатов решения перечисленных задач напрямую зависит от применяемых подходов к представлению и обработке текстов [18]. Таким образом с использованием размеченной обучающей выборки историй болезни строились правила. Для их проверки массив файлов *.ann был выгружен для проведения экспериментов по точности соответствия формальных правил проведенной вручную разметке. Потенциальные факторы риска с указанием количества документов, в которых была произведена их разметка, представлены в табл. 1.

Табл. 1

Термины и количество документов, в которых они были размечены

Термин	Количество документов, в которых он встречался
obesity	14
dizziness	25
headache	28
hypertension	69
atherosclerosis	39
blood pressure	1
max blood pressure	32
usual blood pressure	14

На основе разметки создавалась таблица со строкой для каждого документа, где столбцами были названия рассматриваемых терминов и значениями в строках было значение используемого термина. Такая же таблица создавалась и для результатов работы парсера. По значениям из таблиц для каждого термина рассчитывались показатели точности, полноты, f-меры и «правильности» (accuracy). Их значения можно увидеть в табл. 2.

Как видно из таблицы, F-мера для большинства терминов превышает 90%. Исключение составляют термины, связанные с артериальным давлением, оценка которых не достигает таких же значений. Это может быть вызвано большим количеством разных способов написания показателя артериального давления, которые не были учтены в парсере и небольшим количеством их встречаемости в размеченных текстах, что видно в табл. 1. Последнее может быть обусловлено особенностями анализируемого материала (невозможностью получения необходимой информации ввиду состояния больных или нормальные значения при эффективной медикаментозно управляемой гипертензии.). Стоит также отметить, что такой подход привел к результатам сравнимым с результатами из [19], где использовалось машинное обучение. Это свидетельствует о том, что подход с машинным обучением может достичь результатов обработки человеком.

Заключение

Полученные в процессе исследования результаты позволяют сделать вывод об эффективности и целесообразности использованного подхода для оценки факторов риска на реальных медицинских

Табл. 2

Термины и оценки качества их определения

Термин	Правильность (accuracy)	F-мера (F1)	Точность (Precision)	Полнота (Recall)
obesity	0.979167	0.933333	0.875000	1.000000
dizziness	0.968750	0.938776	0.958333	0.920000
headache	0.947917	0.909091	0.925926	0.892857
hypertension	0.927083	0.951049	0.918919	0.985507
atherosclerosis	0.968750	0.961039	0.973684	0.948718
blood pressure	0.135417	0.023529	0.011905	1.000000
max blood pressure	0.677083	0.060606	1.000000	0.031250
working blood pressure	0.927083	0.666667	1.000000	0.500000

данных. Дальнейшее улучшение результатов оценивания может быть достигнуто при увеличении количества размеченных текстов и уточнении созданных правил. Стоит заметить, что, возможно, использование машинного обучения потребует еще большего количества размеченных текстов для достижения приемлемых результатов, но в этом случае не потребуется создания правил или шаблонов. В целях уменьшения количества необходимых для разметки текстов может использоваться активное обучение.

Литература

1. *Putaala J., Metso A.J., Metso T.M., Konkola N., Kraemer Y., Haapaniemi E., Kaste M., Tatlisumak T.* Analysis of 1008 consecutive patients aged 15 to 49 with first-ever ischemic stroke the Helsinki young stroke registry // Stroke. 2009, vol. 40, no. 4, pp. 1195–1203.
2. *Tibaek M., Dehlendorff C., Jørgensen H.S., Forchhammer H.B., Johnsen S.P., Kammersgaard L.P.* Increasing incidence of hospitalization for stroke and transient ischemic attack in young adults: a registry-based study // Journal of the American Heart Association. 2016, vol. 5, no.5, e003158.
3. *Zhang F-L., Guo Z-N., Wu Y-H., Liu H.-Y., Luo Y., Sun M.-S., Xing Y.-Q., Yang Y.* Prevalence of stroke and associated risk factors: a population based cross sectional study from northeast China // BMJ Open. 2017, vol. 7, no. 9, e015758.
4. *Cesario E., Congiusta A., Talia D., Trunfio P.* Data analysis services in the knowledge grid // Data Mining Techniques in Grid Computing Environments / W. Dubitzky (Ed.). John Wiley & Sons, 2008. Pp.17–36.
5. *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс. 2015. 400 с.
6. *Shelmanov A., Liventsev V., Kireev D., Khromov N., Panchenko A., Fedulova I., Dylov D.V.* Active Learning with Deep Pre-trained Models for Sequence Tagging of Clinical and Biomedical Texts // IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2019, pp. 482-489.
7. *Li X., Feng J., Meng Y., Han Q., Wu F., Li J.* A Unified MRC Framework for Named Entity Recognition. 2019. URL: https://arxiv.org/abs/1910.11476.
8. *Dligach D., Bethard S., Becker L., Miller T., Savova G.K.* Discovering body site and severity modifiers in clinical texts // Journal of the American Medical Informatics Association. 2014, vol. 21, no. 3, pp. 448-454.
9. *Banerjee Ch., Chimowitz M.I.* Stroke Caused by Atherosclerosis of the Major Intracranial Arteries // Circulation Research. 2017, vol. 120, no. 3, pp.502–513.
10. *Price A.J., Wright F.L., Green J., Balkwill A., Kan S.W., Yang T.O., Floud S., Kroll M.E., Simpson R., Sudlow C.L.M., Beral V., Reeves G.K.* Differences in risk factors for 3 types of stroke: UK prospective study and meta-analyses // Neurology. 2018, vol. 90, no. 4, pp. e298-e306.
11. *Murakami K., Asayama K., Satoh M., Inoue R., Tsubota-Utsugi M., Hosaka M., Matsuda A., Nomura K., Murakami T., Kikuya M., Metoki H., Imai Y., Ohkubo T.* Risk Factors for Stroke among Young-Old and Old-Old Community-Dwelling Adults in Japan: The Ohasama Study // Journal of Atherosclerosis and Thrombosis. 2017, vol. 24, no. 3, pp. 290-300.
12. *Lewington S., Clarke R., Qizilbash N., Peto R., Collins R.* Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61

prospective studies // Lancet. 2002, vol. 360, no. 9349, pp. 1903–1913.
13. *Kroll M.E., Green J., Beral V., Sudlow C.L., Brown A., Kirichek O., Price A., Yang T.O., Reeves G.K.* Adiposity and ischemic and hemorrhagic stroke // Neurology. 2016, vol. 87, no. 14, pp. 1473–1481.
14. *Верещагин Н.В.* Недостаточность кровообращения в вертебро-базилярной системе // Consilium Medicum. Головокружение (Приложение). 2001. Т.3. №15. С. 13-18.
15. *Кадыков А.С., Манвелов Л.С., Шахпаронова Н.В.* Хронические сосудистые заболевания головного мозга. Дисциркуляторная энцефалопатия. 4-е изд. М.: ГЭОТАР-Медиа. 2018. 288 с.
16. *Zhang Y., Bai L., Shi M., Lu H., Wu Y., Tu J., Ni J., Wang J., Cao L., Lei P., Ning X.* Features and risk factors of carotid atherosclerosis in a population with high stroke incidence in China // Oncotarget. 2017, vol. 8, no. 34, pp. 57477–57488.
17. *Mitchell A.B., Cole J.W., McArdle P.F., Cheng Y.-Ch., Ryan K.A., Sparks M.J., Mitchell B.D., Kittner S.J.* Obesity Increases Risk of Ischemic Stroke in Young Adults // Stroke. 2015, vol. 46, no. 6, pp. 1690–1692.
18. *Смирнов И.В., Шелманов А.О.* Семантико-синтаксический анализ естественных языков. Часть I. Обзор методов синтаксического и семантического анализа текстов // Искусственный интеллект и принятие решений. 2013. №1. С. 41-54.
19. *Шелманов А.О., Смирнов И.В., Вишнева Е.А.* Извлечение информации из клинических текстов на русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). 2015. №2. С. 560-572.

Благосклонов Николай Алексеевич. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и Управление» РАН, Москва, Россия. Инженер-исследователь. Количество печатных работ: 16. Область научных интересов: системы поддержки принятия решений, инженерия знаний, экспертные системы, интеллектуальные системы, неоднородные семантические сети. E-mail: nblagosklonov@gmail.com

Донитова Виктория Владимировна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и Управление» РАН, Москва, Россия. Научный сотрудник. Количество печатных работ: 12. Область научных интересов: извлечение знаний, интеллектуальные системы, системы поддержки принятия решений, экспертные системы. E-mail: vdonitova@gmail.com

Киреев Данил Алексеевич. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и Управление» РАН, Москва, Россия. Техник 2-ой категории. Количество печатных работ: 1. Область научных интересов: машинное обучение, глубокое обучение, активное обучение, обработка естественного языка, компьютерное зрение, извлечение именованных сущностей, программирование микроконтроллеров. E-mail: kireev@isa.ru

Кобринский Борис Аркадьевич. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и Управление» РАН, Москва, Россия. Заведующий отделом. Доктор медицинских наук, профессор. Количество печатных работ: более 500 (в т.ч. 15 монографий). Область научных интересов: инженерия знаний, экспертные системы, интеллектуальные системы, нечеткие системы, системы поддержки принятия решений. E-mail: kba_05@mail.ru (Ответственный за переписку)

Смирнов Иван Валентинович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и Управление» РАН, Москва, Россия. Заведующий отделом. Кандидат физико-математических наук, доцент. Количество печатных работ: 97. Область научных интересов: интеллектуальный анализ текстов и данных, обработка естественного языка. E-mail: ivs@isa.ru

Linguistic analysis of electronic health records for extraction of stroke risk factors

Blagosklonov N.A., Donitova V.V., Kireev D.A., Kobrinskii B.A., Smirnov I.V.
Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. Identification and assessment of risk factors associated with diseases are necessary to increase the effectiveness of preventive measures. This problem is particularly important for such a socially significant disease as stroke. The use of automated methods for analyzing large arrays of electronic health records can increase the efficiency of extracting information about risk factors. This work presents one of these methods, that is based on using the constructed rules and a linguistic parser..

Keywords: *linguistic parser, text markup, risk factors, stroke.*
DOI: 10.14357/20790279200309

References

1. *Putaala J., Metso A.J., Metso T.M., Konkola N., Kraemer Y., Haapaniemi E., Kaste M. and Tatlisumak T.* 2009. Analysis of 1008 consecutive patients aged 15 to 49 with first-ever ischemic stroke the Helsinki young stroke registry. Stroke 40(4):1195–1203. doi: 10.1161/STROKEAHA.108.529883.
2. *Tibaek M., Dehlendorff C., Jørgensen H.S., Forchhammer H.B., Johnsen S.P. and Kammersgaard L.P.* 2016. Increasing incidence of hospitalization for stroke and transient ischemic attack in young adults: a registry-based study. Journal of the American Heart Association 5(5):e003158. doi: 10.1161/JAHA.115.003158.
3. *Zhang F-L., Guo Z-N., Wu Y-H., Liu H.-Y., Luo Y., Sun M.-S., Xing Y.-Q. and Yang Y.* 2017. Prevalence of stroke and associated risk factors: a population based cross sectional study from northeast China. BMJ Open 7(9):e015758. doi: 10.1136/bmjopen-2016-015758.
4. *Cesario E., Congiusta A., Talia D. and Trunfio P.* 2008. Data analysis services in the knowledge grid. In: W. Dubitzky, ed. Data Mining Techniques in Grid Computing Environments. John Wiley & Sons. pp.17–36. doi: 10.1002/9780470699904.ch2
5. *Flah P.* 2015. Mashinnoe obuchenie. Nauka i iskusstvo postroenija algoritmov, kotorye izvlekajut znanija iz dannyh [Machine learning. Science and art of building algorithms that extract knowledge from data]. Moscow: DMK Press. 400 p.
6. *Shelmanov A., Liventsev V., Kireev D., Khromov N., Panchenko A., Fedulova I. and Dylov D.V.* 2019. Active Learning with Deep Pre-trained Models for Sequence Tagging of Clinical and Biomedical Texts . IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 482-489. doi: 10.1109/BIBM47256.2019.8983157.
7. *Li X., Feng J., Meng Y., Han Q., Wu F. and Li J.* 2019. A Unified MRC Framework for Named Entity Recognition. 2019. Available at: https://arxiv.org/abs/1910.11476 (accessed May 26, 2020).
8. *Dligach D., Bethard S., Becker L., Miller T. and Savova G.K.* 2014. Discovering body site and severity modifiers in clinical texts. Journal of the American Medical Informatics Association 21(3):448-454. doi: 10.1136/amiajnl-2013-001766.
9. *Banerjee Ch. and Chimowitz M.I.* 2017. Stroke Caused by Atherosclerosis of the Major Intracranial Arteries. Circulation Research 120(3):502–513. doi: 10.1161/CIRCRESAHA.116.308441.
10. *Price A.J., Wright F.L., Green J., Balkwill A., Kan S.W., Yang T.O., Floud S., Kroll M.E., Simpson R., Sudlow C.L.M., Beral V. and Reeves G.K.* 2018. Differences in risk factors for 3 types of stroke: UK prospective study and meta-analyses. Neurology 90(4):e298-e306. doi: 10.1212/WNL.0000000000004856.
11. *Murakami K., Asayama K., Satoh M., Inoue R., Tsubota-Utsugi M., Hosaka M., Matsuda A., Nomura K., Murakami T., Kikuya M., Metoki H., Imai Y. and Ohkubo T.* 2017. Risk Factors for Stroke among Young-Old and Old-Old Community-Dwelling Adults in Japan: The Ohasama Study. Journal of Atherosclerosis and Thrombosis 24(3):290-300. doi: 10.5551/jat.35766.
12. *Lewington S., Clarke R., Qizilbash N., Peto R. and Collins R.* 2002. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. Lancet 360(9349):1903–1913. doi: 10.1016/S0140-6736(02)11911-8.
13. *Kroll M.E., Green J., Beral V., Sudlow C.L., Brown A., Kirichek O., Price A., Yang T.O. and Reeves G.K.* 2016. Adiposity and ischemic and hemorrhagic stroke. Neurology 87(14):1473–1481. doi: 10.1212/WNL.0000000000003171.
14. *Vereshhagin N.V.* 2001. Nedostatochnost' krovoobrashhenija v vertebro-baziljarnoj sisteme [Circulatory failure in the vertebro-basilar system]. Consilium Medicum. Golovokruzhenie (Prilozhenie) [Consilium Medicum. Dizziness (Appendix)] 3(15):13-18.

15. *Kadykov A.S., Manvelov L.S. and Shahparonova N.V.* 2018. Hronicheskie sosudistye zabolevanija golovnogo mozga. Discirkuljatornaja encefalopatija. 4-e izd [Chronic vascular diseases of the brain. Encephalopathy. 4th ed.]. Moscow: GEOTAR-Media. 288 p.
16. *Zhang Y., Bai L., Shi M., Lu H., Wu Y., Tu J., Ni J., Wang J., Cao L., Lei P. and Ning X.* 2017. Features and risk factors of carotid atherosclerosis in a population with high stroke incidence in China. Oncotarget 8(34):57477–57488. doi: 10.18632/oncotarget.15415.
17. *Mitchell A.B., Cole J.W., McArdle P.F., Cheng, Y.-Ch., Ryan, K.A., Sparks, M.J., Mitchell, B.D. and Kittner, S.J.* 2015. Obesity Increases Risk of Ischemic Stroke in Young Adults. Stroke 46(6):1690–1692. doi: 10.1161/STROKEAHA.115.008940.
18. *Smirnov I.V. and Shelmanov A.O.* 2013. Semantiko-sintaksicheskij analiz estestvennyh jazykov. Chast' I. Obzor metodov sintaksicheskogo i semanticheskogo analiza tekstov [Semantic-syntactic analysis of natural languages. Part I. A review of methods for semantic and syntactic analysis of text]. Iskusstvennyj intellekt i prinjatie reshenij [Artificial Intelligence and Decision Making] (1):41-54.
19. *Шелманов А.О., Смирнов И.В., Вишнева Е.А.* Извлечение информации из клинических текстов на русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). 2015. №2. С. 560-572. Shelmanov, A.O., Smirnov, I.V. and Vishneva, E.A. 2015. Information extraction from clinical texts in russian. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue". Volume 1 of 2:560-572.

N.A. Blagosklonov Place of work: Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Vavilova str. 44, kor.2, Moscow, 119333, Russian Federation, E-mail: nblagosklonov@gmail.com

V.V. Donitova Place of work: Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Vavilova str. 44, kor.2, Moscow, 119333, Russian Federation. E-mail: vdonitova@gmail.com

D.A. Kireev Place of work: Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Vavilova str. 44, kor.2, Moscow, 119333, Russian Federation. E-mail: kireev@isa.ru

B.A. Kobrinskii PhD, Professor. Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Vavilova str. 44, kor.2, Moscow, 119333, Russian Federation. E-mail: kba_05@mail.ru

I.V. Smirnov PhD, Assoc. Professor. Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Vavilova str. 44, kor.2, Moscow, 119333, Russian Federation. E-mail: ivs@isa.ru