И. А. Коломойцева
Донецкий национальный технический универсиет
(Украина, 83000, Донецк, ул. Артема, 58,
тел.(062) 3010729, E-mail: kolomoit@r5.dgtu.donetsk.ua)

Функциональная модель медицинского естественно-языкового текста

Введение. В последнее время все больше и больше срециалистов в различных областях обращаются при поиске информации к Интернету. Огромное количество разрозненной и во многих случаях повторяющейся информации требует автоматизированной обработки.

В настоящее время технологии полного и точного автоматического анализа произвольного текста пока не существует. Наименее разработанными являются модели и методы семантического уровня [1].

Области применения семантического анализа очень разнообразны [1]. Для данной статьи актуальной является задача перехода от плохо структурированной (ЕЯ-текст) к хорошо структурированной информации, которую можно обработать стандартными и высокоэффективными средствами информационных технологий.

Методы и средства семантического анализа можно разделить на два направления [1]:

1) средства формализации фактологической информации (для СУБД);

2) средства формализации номологической информации (для экспертных систем).

Именно плохая структурированность медицинского ЕЯ-текста существенно осложняет его обработку. А потребность анализа больших объемов текстологических данных есть. В первую очередь, это относится к современному, быстро развивающемуся разделу медицины – сравнительной медицине. В связи с этим построение алгоритма извлечения фактологической информации (семантического анализа) из медицинского ЕЯ-текста и построения БД является актуальной теоретической и практической задачей.

Основной задачей этой работы является разработка методики анализа связного текста для работы с медицинскими естественно-языковыми текстами (описания лекарств) с использованием семантико-синтаксического анализатора и базы знаний конкретной предметной области, основанного на функциональном представлении естественно-языкового текста. В качестве основы был использован семантический словарь В. А. Тузова.

Общая характеристика моделей. Среди наиболее известных и влиятельных работ, посвящённых формальному описанию языков, можно выделить теорию формальных грамматик Н. Хомского [3] и модель «смысл ↔ текст» И. Мельчука [3]. Идеи Хомского лежат в основе известных алгоритмов анализа текстов компьютерных программ. Многие синтаксические анализаторы англоязычных текстов так или иначе используют грамматики Хомского. Модель Мельчука изначально предназначалась для изучения проблемы формализации естественных языков. Кроме того, модель «смысл ↔ текст» не обладает достаточной языковой независимостью (т.е. ориентирована на тексты на славянских языках), что препятствует её распространению на Западе.

Данная работа базируется на модели формализации русского языка, предложенной проф. В. Тузовым [4]. В настоящее время подход В. Тузова ещё не получил широкого распространения, но программные продукты, его использующие, уже существуют.

Описание семантического анализатора. В статье предлагается подход, в основе которого лежит компьютерное толкование смысла слова на формальном семантическом языке и функциональное использование этого толкования при вычислении смысла предложения.

С точки зрения этого подхода любое слово русского языка является именем (названием) функции f(x1, …, xn), которая связывается с этим словом и называется его семантикой. Каждое свое конкретное значение слово получает только после подстановки конкретных значений. Смысл слова вычисляется в процессе выполнения функции f.

Предложение – единая законченная суперпозиция функций. Смысл предложения вычисляется в процессе построения и выполнения суперпозиции [4].

Семантический анализатор в процессе построения суперпозиции выполняет два основных действия:

1) выбор правильного значения (компьютерного толкования) слова;

2) связывание выбранных значений в осмысленные выражения (целостные словосочетания), т.е. в выражения, которые имеют независимое семантическое толкование.

Часто используемое в лингвистике понятие валентность следует понимать буквально, в химическом смысле. С точки зрения информатики, присоединяемые слова являются аргументами, из которых присоединяющее их слово строит новую конструкцию, семантика которой может существенно отличаться от семантики ее составляющих. Любой достаточно развитый язык имеет функциональную природу, и только суперпозиция функций адекватна структуре предложений такого языка.

Функциональная природа языка проявляется на всех его уровнях – от механизма словообразования до механизма построения текста. Достаточно посмотреть на аффикс как на функцию, аргументом которой является корень слова, чтобы увидеть удивительную регулярность механизма словообразования русского языка. Благодаря этой регулярности вначале удалось формализовать семантику словообразования, затем автоматически построить семантическое описание большого количества производных слов, сведя эти описания к описанию слов более простых по морфемному составу. Это позволило существенно автоматизировать процесс построения компьютерного семантического словаря и, в конечном счете, построить его [4].

Качество любого семантического анализатора можно оценивать по тому, как он вычисляет семантико-грамматическое значение предложно-падежных форм. Точность семантического анализа прямо зависит от качества и полноты семантического словаря.

Процесс анализа текста разбивается на три составные части: морфологический анализ, предварительная пословная обработка текста и семантический анализ.

Морфологический анализатор осуществляет обработку текста, вычисляя необходимые для дальнейшего анализа морфологические (грамматические) характеристики каждой словоформы. На этапе предварительной обработки для

каждой словоформы вычисляются наборы ее морфосемантических значений. На этапе семантического анализа осуществляется выбор конкретного морфосемантического значения словоформы и связывание всех слов предложения в единую семантическую структуру. Морфологический анализатор использует морфологический (грамматический) словарь, на двух последующих этапах используется семантический словарь.

Словарная статья компьютерного семантического словаря содержит заголовочное слово и его толкование на семантическом языке. Многие слова содержат более одного толкования. Основная задача семантического анализатора при анализе конкретного предложения – правильный выбор альтернативы. Этот выбор определяется классом и предложно-падежной формой аргументов, которые способна присоединить к себе та или иная альтернатива.

Постановка задачи анализа текста. Задача анализа текста имеет два варианта: задача понимания текста и задача извлечения смыслов.

Задача понимания текста означает полное и однозначное сопоставление (обычно небольшого) фрагмента текста некоторой формальной структуре, описывающей его смысл. Это на практике проецируется на перевод или на диалог с пользователем.

Задача извлечения смыслов ставит целью выборку из текста всех элементов понимания, полных и частичных, при этом допускается противоречивость элементов между собой. Эта задача направлена на обработку больших массивов текстов, в частности на поиск, фильтрацию, статистическую обработку.

Для первой задачи различные варианты прочтения многозначных элементов текста анализируются последовательно, и первое адекватное прочтение всего фрагмента прекращает анализ. При этом если частичное понимание фрагмента не складывается в общую структуру, оно само по себе бесполезно, и требуется удалить это частичное понимание, чтобы гарантировать его неучастие в анализе альтернативных вариантов прочтения. Таким образом, первая задача приводит к применению поиска с возвратами (бэктрекинга) [4].

В случае использования второго варианта рассматриваются все смыслы, полные и частичные, извлекаемые из текста, и в том числе альтернативные прочтения одного и того же [4].

В данной работе будет рассмотрен второй случай анализа текста, использующий анализатор, предложенный В.А.Тузовым [4].

Для того чтобы получить из общего текста нужную информацию, её необходимо идентифицировать. Для этого понадобиться структура, в которую помещается ключевое слово в именительном падеже (если это существительное или прилагательное) или в инфинитиве (если это глагол) и номер. Данная структура необходима для поиска ключевого слова в семантическом словаре, а номер – для определения значения, в котором используется это слово.

Точность обработки предложений будет зависеть от **полноты семантического словаря**, который при необходимости можно дополнять. Кроме того, слова, указанные в структуре, обязательно должны присутствовать в тексте, иначе текст будет распознан не верно.

Рассмотрим работу распознавания текста на примере. Имеется структурированный текст по описанию лекарства, приведенный на рисунке 1. Задача: получить только интересующую нас информацию. Например, необходимо узнать, через какое количество времени начинает действовать препарат.

Анальгин

Лекарственные формы

субстанция, таблетки 500мг, раствор для инъекций, раствор для инъекций 1г, раствор для инъекций 25%, раствор для инъекций 50%, раствор 25%, раствор 50%, капсулы 250мг, таблетки для детей 50мг, свечи для детей 100мг, свечи для детей 250мг, раствор для инъекций 2.5г, таблетки 40мг, субстанция 200г, субстанция 500г, субстанция 1кг, субстанция 25кг

Производители

Аганал Трейдерс(Индия), АЙ СИ ЭН Лексредства(Россия), АЙ СИ ЭН Марбиофарм(Россия), АЙ СИ ЭН Октябрь(Россия), Акрихин ХФК(Россия), Алкалоид(Македония), Аллерген Ставрополь(Россия), Алтайвитамины(Россия), Аналитико-менеджерская группа(Россия), Антивирал(Россия), Асфарма(Россия), Белвитамины(Россия), Белвитамины-Время Фарм.Произв.Компания(Россия), Белгородвитамины(Россия), Белмедпрепараты(Беларусь), Берлин-Хеми АГ(Германия), Биомед(Россия)

ФармГруппа

Анальгетики-антипиретики - производные пиразолона

Состав

Действующее вещество - Метамизол натрия.

Фармакологическое действие

Противовоспалительное, анальгезирующее, жаропонижающее. Угнетает активность циклооксигеназы, снижает образование эндоперекисей, брадикининов, некоторых простагландинов, свободных радикалов, ингибирует перекисное окисление липидов. Препятствует проведению болевых экстра- и проприоцептивных импульсов по пучкам Голля и Бурдаха, повышает порог возбудимости таламических центров болевой чувствительности, увеличивает теплоотдачу. При приеме внутрь быстро и полно абсорбируется. Разрушается в печени. Экскреция проходит через почки. Действие развивается через 20-40 минут.

Показания к применению

Артралгии, ревматизм, хорея, боли: головная, зубная, менструальная, невралгия, ишиалгия, миалгия, при коликах (почечная, печеночная, кишечная), инфаркте легкого, инфаркте миокарда, расслаивающей аневризме аорты, тромбозе магистральных сосудов, воспалительных процессах (плеврит, пневмония, люмбаго, миокардит), травмах, ожогах, декомпрессионной болезни, опоясывающем лишае, опухолях, орхите, панкреатите, перитоните, перфорации пищевода, пневмотораксе, посттрансфузионных осложнениях, приапизме; лихорадочный синдром при острых инфекционных, гнойных и урологических заболеваниях (простатит), укусах насекомых (комары, пчелы, оводы и др.).

Противопоказания

Гиперчувствительность, угнетение кроветворения (агранулоцитоз, цитостатическая или инфекционная нейтропения), тяжелые нарушения функции печени или почек, простагландиновая бронхиальная астма, наследственная гемолитическая анемия, связанная с дефицитом глюкозо-6-фосфатдегидрогеназы, беременность, кормление грудью (на время лечения прекращают).

Побочное действие

Гранулоцитопения, агранулоцитоз, тромбоцитопения, геморрагии, гипотония, интерстициальный нефрит, аллергические реакции (в т.ч. синдромы Стивенса - Джонсона, Лайелла, бронхоспазм, анафилактический шок).

Передозировка

Симптомы: гипотермия, выраженная гипотензия, сердцебиение, одышка, шум в ушах, тошнота, рвота, слабость, сонливость, бред, нарушения сознания, судорожный синдром; возможно развитие острого агранулоцитоза, геморрагического синдрома, острой почечной и печеночной недостаточности. Лечение: индукция рвоты, чреззондовое промывание желудка, назначение солевых слабительных, активированного угля и проведение форсированного диуреза, ощелачивание крови, симптоматическая терапия, направленная на поддержание жизненно важных функций.

Особые указания

Необходим врачебный контроль. Не рекомендуется регулярный длительный прием вследствие миелотоксичности. Исключается использование для снятия острых болей в животе (до выяснения причины). При назначении больным с острой сердечно-сосудистой патологией необходим тщательный контроль за гемодинамикой. С осторожностью применяют у пациентов с уровнем систолического артериального давления ниже 100 мм рт.ст., с анамнестическими указаниями на заболевания почек (пиелонефрит, гломерулонефрит) и при длительном алкогольном анамнезе. При применении метамизола возможно красное окрашивание мочи за счет выделения метаболита [8].

Рисунок 1 – Описание лекарства «Анальгин»

В структуру помещаем слово ДЕЙСТВИЕ. Смысловые понятия этого слова в семантическом словаре приведены на рисунке 2.

1 - работа, деятельность, совершение чего-либо. Скоординированные действия;
2 - совокупность поступков кого-нибудь. Непродуманные действия;
3 - работа, функционирование какой-нибудь машины, агрегата, предприятия и т.д.;
4 - влияние на кого-, что-нибудь. Позитивное действие;
5 - совокупность и развитие действий в литературных произведениях, кино и т.д.;
6 - боевые действия;
7 - законченная часть драматического, оперного, балетного произведения (синоним акта);
8 - основной вид математического произведения.

Рисунок 2 – Смысловые понятия слова «ДЕЙСТВИЕ» в семантическом словаре

Действие
…
4. ДЕЙСТВИЕ{СущНеодуш$11227~@ОНО$5@Им$11227~@ОНО$5@Вин} $15(!Им!Вин,!Что)
…
6. ДЕЙСТВИЕ{СущНеодуш$11251~@ОНО$5@Им} $11251(!Им,!Какое)

Развивается (имеет два смысла: проявлять новые способности и подвергаться действию раскручивания)
1. РАЗВИВАТЬСЯ{Глаг} Oper00(!Глаг,Magn_a~СПОСОБНОСТЬ$1103(!Им))
2. РАЗВИВАТЬСЯ{Глаг} Caus(ПРИЧИНА$10/05~!От,Lab(!Глаг,РАСКРУЧИВАНИЕ$125~!Им))
Через
1.ЧЕРЕЗ{Предл@черезВин@черезПред@черезКого@вОНИ$5@Им}$71(!Вин\!Пред\!Кого\!ОНИ$5@Им)
2.ЧЕРЕЗ {Предл $13~@Как} Direkt_y(#,СПОСОБНОСТЬ 14~!Вин) ,
3.ЧЕРЕЗ {Предл $15~@Что} Direkt_y(#,ВРЕМЯ$3~!Вин)
Symbol: 20-40
Минут
1.МИНУТА{СущНеодуш$3~@ОНА$5@Вин} $327(!поРод,!вВин\!наВин\!черезВин)
Symbol: .

Рисунок 3 – Альтернативы пословной обработки предложений

Для данного примера выбирается вариант под номером 4, поэтому в структуру, кроме слова, заносится еще и цифра 4.

Семантический анализатор начнет поиск слова ДЕЙСТВИЕ в каждом предложении, начиная с первого. По тексту это слово впервые встречается в предложении: «Действие развивается через 20-40 минут». Это предложение

первым начинает анализироваться. Все возможные альтернативы пословной обработки приведены на рисунке 3.

На следующем этапе семантического анализа происходит отбор нужных альтернатив. В данном примере морфологические признаки слова МИНУТА позволят однозначно определить альтернативу предлога ЧЕРЕЗ. Совпадение класса и падежа указывают на вариант под номером 3. Морфологическая часть предлога образует класс под номером 15 с вопросом ЧТО – это совпадает с 4-ой семантикой слова ДЕЙСТВИЕ. Далее аналогичным образом по падежу и классу определяется однозначность слова РАЗВИВАТЬСЯ, что соответствует первому варианту.

После этого можно делать проверку чисел на совпадение. В структуре слову ДЕЙСТВИЕ соответствовала цифра 4. Семантическим анализатором для этого слова была выбрана альтернатива под номером 4. Цифры совпадают, значит предложение подходит и оно будет записано в выходной файл в качестве результата. Таким образом, после заполнения структуры словами нужного смысла получаются необходимые смысловые предложения.

Выводы. Модель семантического анализатора Тузова В.А. применима к медицинским текстам с изменениями, учитывающими особенности предметной области.

Эффективно работающая модель анализа связного текста требует:

– полной классификации слов медицинских естественно-языковых текстов (так как от точности и полноты классификации зависит структура и работа следующих этапов анализа текста);

– модуля анализа и пополнения семантического словаря новыми альтернативами (так как именно от полноты словаря будет зависеть эффективность и точность анализа текста).

Анализ текста с помощью данного семантического анализатора дает неплохие результаты. Однако, как и любой другой метод, он имеет свои преимущества и недостатки. К недостаткам можно отнести жесткие правила, необходимые для предложенного метода (точный выбор альтернативы, полнота семантического словаря), нарушение которых может привести к неправильным результатам определения информации. Большим преимуществом данного подхода является «умный» поиск информации, а не просто подбор слов по совпадению. Это дает возможность оперировать наиболее точной информацией, тем более учитывая жизненную важность предметной области и количество информации, существующей на сегодняшний день.

1. *Рубашкин В.Ш.* Семантический компонент в системах понимания текста // Труды Десятой национальной конференции по искусственному интеллекту с международным участием (КИИ-2006). - М.: Физматлит. - 2006. - Т. 2. - С. 455-463.

2. *Хорошевский В.Ф.* Оценка систем извлечения информации из текстов на естественном языке: кто виноват, что делать // Труды Десятой национальной конференции по искусственному интеллекту с международным участием (КИИ-2006). - М.: Физматлит. - 2006. - Т. 2. - С. 464-478.

3. *Тузов В.А.* Компьютерная семантика русского языка. – СПб.: Изд-во СПбГУ, 2004. – 400 с.

4. *Тузов В.А.* Компьютерная семантика русского языка //Труды международной конференции «Диалог 2001».

5. *Поспелов Д. А*. Логико-лингвистические модели в системах управления. – М.: Энергоиздат, 1981. – 232 с.5.