Принципы формального отображения семантики лексических единиц, предложений и дискурсов в интеллектуальной поисковой системе Medsearch

И. В. Люстиг, В. А. Фомичев
МИЭМ, 109028 Москва
lyustig@mail.ru; vdrfom@aha.ru

Разрабатываемая интеллектуальная система MEDSEARCH предназначена для получения справочной информации о лекарственных препаратах: общих сведений о препарате, списка лекарственных препаратов, предназначенных для лечения данного заболевания, выявленных побочных эффектов от применения препарата, возможных побочных эффектов от применения препарата с учетом особенностей пациента. Поиск осуществляется в базе данных (БД) описаний лекарственных препаратов, имеющейся в системе. Для поддержки БД в актуальном состоянии осуществляется ее обновление через сеть Интернет путем получения требуемой информации с сайтов фармацевтических компаний. Ведется не обычный поиск по ключевым словам, а семантико-синтаксический анализ электронных документов с целью выявления тех, которые по смыслу соответствуют запросу пользователя. В качестве системы управления БД (СУБД) системы MEDSEARCH выбрана СУБД Oracle.
Описания лекарственных препаратов имеют достаточно строгую структуру, которая подвержена незначительным изменениям, что позволяет применять для смыслового анализа семантико-синтаксические шаблоны фраз и более крупных фрагментов текста. В такие шаблоны могут включаться как конкретные слова, так и обозначения понятий. С каждым семантико-синтаксическим шаблоном связано обозначение выражаемого им смысла. Основное внимание в работе уделяется принципам описания семантических структур при проектировании лингвистического процессора системы MEDSEARCH. Методологической основой для описания семантики лексических единиц, предложений и дискурсов, а также для построения семантико-синтаксических шаблонов является теория стандартных К-языков, предложенная В. А. Фомичевым («Информационные технологии», 2002, № 10, № 11). Эта теория описывает 10 операций на концептуальных структурах, позволяющих строить семантические представления как предложений, так и сложных связных текстов на естественных языках, в том числе и текстов медицинской тематики.

1. Введение

В настоящее время всемирная сеть Интернет предоставляет доступ к огромному числу ресурсов из всевозможных отраслей знания, в том числе, к ресурсам медицинской тематики. Достижения в области медицины, например, появление новых лекарственных препаратов, находят отражение на сайтах фармацевтических компаний. Но полная

смысловая обработка всей доступной в сети Интернет справочной информации по интересующей тематике, как правило, требует существенных временных затрат, т. к. изобилие доступных ресурсов препятствует быстрому и легкому обнаружению пользователем-медиком информации в контексте стоящей перед ним проблемы, поскольку поиск в сети осуществляется, главным образом, по ключевым словам.
Разрабатываемая система MEDSEARCH [1] предназначена для облегчения и ускорения получения справочной информации о лекарственных препаратах по следующим типам запросов: общие сведения о препарате, перечень лекарств, применяемых для лечения данного заболевания, список побочных эффектов от применения препарата, характер возможных побочных эффектов с учетом особенностей пациента.
Потенциальными пользователями системы являются медицинские работники, пациенты и их родственники, а так же сотрудники фармацевтических компаний, занимающиеся исследованием рынка.
Система MEDSEARCH имеет следующие составляющие: базу данных (БД) описаний лекарственных препаратов, лингвистическую БД, базу знаний о медицине, компоненты, семантико-синтаксический анализатор текстовых полей электронных документов, и модули взаимодействия с пользователем и с сетью Интернет для получения новых данных в БД. Поиск информации осуществляется путем анализа существующих электронных описаний препаратов, хранящихся в БД, т. е. только локально хранящихся данных.
Практика показывает, что поиск по ключевым словам, применяемый в большинстве поисковых машин в сети Интернет, не дает высококачественного результата: число найденных документов велико, а их релевантность низка. Для повышения качества поиска в системе MEDSEARCH применяется семантико-синтаксический анализ документов, позволяющий находить тексты и извлекать из них фрагменты, выражающие определенный смысл. Описания лекарственных препаратов имеют достаточно строгую структуру, поэтому они поддаются формальному описанию с помощью специальных семантико-синтаксических шаблонов предложений или более крупных фрагментов текста.
За рубежом до последнего времени для формального представления смысла текстов на естественных языках (ЕЯ) использовались преимущественно средства, предоставляемые теорией семантических сетей, теорией концептуальных графов, эпизодической логикой, теорией представления дискурсов. В нашей стране применялись различные варианты языков, предлагаемых теорией семантических сетей, теорией фреймоподобных языков представления знаний, теорией расширенных семантических сетей и компьютерной семантикой есттественного языка. Однако все перечисленные теоретические подходы не дают средств описания смысловой структуры произвольных ЕЯ-текстов, встречающихся в реальных предметных областях. Между тем, проблема автоматизации смыслового анализа медицинских текстов требует формального аппарата для построения семантических представлений (СП) текстов, удобного для создания эффективных анализаторов ЕЯ-текстов. Поэтому в качестве методологической основы для построения СП ЕЯ-текстов и семантико-синтаксических шаблонов выбрана теория стандартных К-языков (СК-языков), предложенная В. А. Фомичевым [2-4]. Эта теория предоставляет принципиально новую математическую модель для описания структурированных значений предложений и связных ЕЯ-текстов. Она задает 10 операций на концептуальных структурах, с помощью которых можно строить СП, по-видимому, сколь угодно сложных естественно-языковых предложений и дискурсов, в том числе текстов медицинской тематики. При этом переход к обработке более широких подъязыков ЕЯ не требует отказа от использовавшихся ранее средств представления смысла текстов.

2. Краткая характеристика принципов построения семантических представлений текстов в теории СК-языков

Каждый стандартный К-язык (СК-язык) определяется некоторым формальным объектом В, называемым концептуальным базисом (к.б.) и являющимся упорядоченной тройкой вида *(S, Ct, Ql)* , где *S, Ct, Ql* – упорядоченные наборы, компонентами которых могут быть множества цепочек, выделенные элементы таких множеств, бинарные отношения на множествах, отображения на множествах. Следуя [3, 4], охарактеризуем более подробно компонент *S*, называемый сортовой системой (с.с.). Каждая с.с. является упорядоченной четверкой вида *(St, P, Gen, Tol)* , компоненты которого интерпретируются следующим образом. *St* является конечным множеством символов, называемых сортами и обозначающих наиболее общие понятия рассматриваемой предметной области (ПО): пространственный объект, интеллектуальная система и т. д. Каждое такое понятие характеризует сущность, не рассматриваемую как упорядоченный набор других сущностей или как множество. Выделяется некоторый сорт *P*, который будет связываться с семантическими представлениями (СП) ЕЯ-текстов, выражающих отдельные высказывания либо являющихся связными повествовательными текстами. *P* называется сортом «смысл сообщения». С помощью бинарного отношения *Gen* на *St* задается иерархия понятий на множестве сортов *St*, т. е. выделяется некоторое подмножество $Gen \subset St \times St$. Например, может выполняться соотношение *(цел. число, нат. число), (вещ. число, цел. число), (простр. объект, физ. об)* $\in Gen$.

Так как многие объекты могут быть охарактеризованы с разных точек зрения, у них есть «координаты» по разным «семантическим осям». Например, с конкретным пациентом можно связать «семантические координаты» «пространственный объект» и «интеллектуальная система». Учитывая это, вводится бинарное отношение совместимости (толерантности) *Tol* на множестве *St*, интерпретируемое следующим образом: если $(s, u) \in Tol \subset St \times St$, то существует такая сущность *x* в рассматриваемой ПО, что с *x* можно связать сорт *s* по одной семантической оси и сорт *u* по другой оси, причем сорт *s* и сорт *u* не являются сравнимыми для отношения *Gen*.

Существование отношения *Tol* учтено при разработке лингвистической базы данных (ЛБД) следующим образом. Важной составляющей ЛБД является лексико-семантический словарь (ЛСС), содержащий информацию о взаимосвязях лексических единиц и соответствующих им семантических единицах. Каждой лексической единице ставится в соответствие один или несколько сортов, причем каждая пара таких сортов связана отношением совместимости.

Определение концептуального базиса, а также группа специальных правил *P[0* , *P[1]*, ..., *P[10]*, которые кратко описываются ниже, дают возможность строить формулы, удобные для описания структурированных значений ЕЯ-текстов. Часть таких формул образует стандартный К-язык (СК-язык), порождаемый рассматриваемым концептуальным базисом. Выражения СК-языков будем называть К-цепочками.

Правило *P[0]* дает начальный запас выводимых формул. Правило *P[1]* предназначено для присоединения информационных единиц, соответствующих словам «некоторый», «каждый», «какой-нибудь», «все», «несколько», «большинство» и т. п. к простым или составным обозначениям понятий. Поэтому *P[1]* позволяет строить формальные аналоги выражений «некоторый человек», «все пациенты», «большинство людей», «некоторый человек ростом 175 см», «все тридцатилетние люди», «все заболевания желудка».

Правило *P[2]* предназначено для построения цепочек вида $f(a_1, ..., a_n)$, где f — обозначение функции, $n \geq 1$, $a_1, ..., a_n$ — формулы. Так, после применения *P[2]* на последнем шаге вывода можно получить цепочки *Эффект(Аспирин), Колич-элем(Эффект*

(Аспирин)). Правило *P[3]* позволяет строить цепочки вида ($a_1 \equiv a_2$), где a_1, a_2 —формулы, полученные при помощи любых правил из *P[0], …, P[10],* и a_1, a_2 обозначают сущности, являющиеся однородными в некотором смысле. Пример К-цепочки для *P[3]* как последнего примененного правила: $y_1 \equiv$ *нек лекарство*(Название, Аспирин).*

Правило *P[4]* позволяет строить К-цепочки вида r (a_1, …, a_n), где r —обозначение n-арного отношения, $n \geq 1$, a_1, …, a_n — К-цепочки. Примеры:

Принадлеж(насморк, Симптомы(простуда)), Подмнож(Симптомы(простуда), Симптомы(грипп)).

Правило *P[5]* предназначено для построения К-цепочек вида $d : v$, где d —К-цепочка, не включающая v, v —переменная, и выполнены некоторые условия. При помощи *P[5]* можно помечать переменными в семантических представлениях текстов на ЕЯ: 1) описания различных сущностей, встречающихся в тексте, 2) СП предложений или более крупных фрагментов текста, на которые имеется ссылка в любой части текста. Примерами К-цепочек для правила *P[5*], примененного на последнем шаге вывода, являются выражения *все чел S1, Меньше(Возраст(П. Сомов), <30, год>) : P1.*

Правило *P[6*] позволяет строить К-цепочки вида $\neg d$, где d — К-цепочка, удовлетворяющая ряду условий. Примеры К-цепочек для *P[6]*: *¬таблетка, ¬Принадлеж(Аспирин, Лекрства(Астма))*. Здесь «¬» — связка «не».

При помощи правила *P[7]* можно строить К-цепочки вида ($a_1 \wedge \ldots \wedge a_n$) или ($a_1 \vee \ldots \vee a_n$), где $n > 1$, a_1, …, a_n — К-цепочки, обозначающие однородные в некотором смысле сущности. В частности, a_1,…,a_n могут быть СП высказываний, описаниями физических объектов, описаниями множеств, состоящих из объектов одной природы, описаниями понятий. Следующие выражения являются примерами К-цепочек для *P[7]*:

(Аспирин ∨ Цитромон ∨ Анальгин),

(Принадлеж((астма ∧ клаустрофобия), Болезни(Человек))

∧ ¬Принадлеж(Аспирин, Лекарства ((СПИД ∨ клаустрофобия ∨ астма)))).

Правило *P[8]* позволяет строить, в частности, К-цепочки вида c*(r_1, b_1), …, (r_n, b_n), где c — информационная единица, обозначающая понятие, для $i = 1, \ldots, n$, r_i —функция с одним аргументом или бинарное отношение, b_i обозначает возможное значение r_i для объектов, характеризующихся понятием c. Например, после применения на последнем шаге вывода правила Р[8] можно получить К-цепочки *лекарство* (Форма, порошок).*

Правило *P*[9] дает возможность строить, в частности, К-цепочки вида *∀v (concept) D* и *∃v(concept) D*, где ∀ — квантор всеобщности, ∃ — квантор существования, *concept* обозначает понятие («человек», «лекарство», «целое число» и др.) или составное понятие («целое число, большее 200» и др.). *D* можно интерпретировать как СП высказывания с переменной v о любой сущности, характеризуемой понятием concept. Примеры:

∀x1(пациент) ∃x2(чел) Моложе(x2, x1), ∃y(болезнь (Носитель, человек)) Больше(Инкуб-период(y), <15, день>).*

Правило *P[10]* позволяет строить, в частности, К-цепочки вида <a_1, …, a_n>, где $n > 1$, a_1, …, a_n —К-цепочки. Выражения вида <a_1, …, a_n> интерпретируются как обозначения n-мерных наборов. Компонентами такого набора могут быть не только обозначения чисел, объектов, но и СП выражений, обозначения множеств, понятий и др. Используя правила *P[10]* и *P[4]* , можно построить цепочку *Лечиться(<Пациент, нек чел * (Имя, 'Петр')>, <Больница, ЦКБ>, <Начал. момент, 2003>),* где *Пациент, Больница, Начал. Момент* — обозначения тематических ролей, т. е. обозначения отношений между значением глагола «лечиться» и значениями зависящих от него в предложении групп слов.

Рассмотренные выше краткие описания правил *P[0* , *P[1]*, ..., *P[10]* дают только самое общее представление об этих правилах. Полные системы математических определений можно найти в работах [2-4].

Разработанный математический подход открывает много новых возможностей построения семантических представлений (СП) ЕЯ-текстов.

Пример 1. Пусть Т1 – относящийся к биологии и медицине дискурс “Все гранулоциты являются полиморфонуклеарными. Это означает, что их ядра многодольны”. Тогда дискурсу Т1 можно поставить в соответствие следующую К-цепочку *Expr1*, интерпретирукмую как СП текста Т1 (т.е. К-представление Т1)

(Свойство (произвольн гранулоцит : x1 , полиморфонуклеарный) : P1) ∧ Пояснение (P1, Следует-из (Ситуация (e1, обладание1 (Агент1, x1)(Объект1, произвольн ядро : x2)), Свойство (x2, многодольный)))) .*

Ключевую роль в построении К-представления *Expr1* сыграло правило Р[5], позволившее ввести метку *x1* для обозначения произвольного гранулоцита, метку *x2* для обозначения произвольного ядра клетки, и метку *P1* для обозначения семантического представления первого предложения из дискурса Т2. Метка *P1* позволяет в структуре СП текста Т1 эксплицировать ссылку на смысл первого предложения текста, даваемую сочетанием “Это означает”.

Пример 2. Пусть Т2 = «Сфигмоманометр — прибор, предназначенный для изменения кровяного давления», тогда Т2 может иметь следующее КП:

*(сфигмоманометр ≡ прибор * (Назначение, измерение1 * (Парам, кровяное-давление)(Субъект, произв чел))) .*

Семантическая единица *Назначение* в этом КП обозначает бинарное отношение. Если пара (А, В) принадлежит этому отношению, то А является физическим объектом, а В - формальным семантическим аналогом выражения, описывающего назначение этого физического объекта.

Пример 3. Пусть Т3 — определение «Тромбин — это фермент, который помогает преобразовать фибриноген в фибрин во время коагуляции». Тогда следующая К-цепочка является возможным КП Т3:

*(тромбин ≡ фермент * (Назначение, оказание-помоши * (Действие, преобразование1* (Исх-объект, нек фибриноген)(Результат1, нек фибрин)(Процесс, нек коагуляция)))).*

3. Средства реализации подхода

В качестве системы управления БД (СУБД) описаний лекарственных препаратов выбрана СУБД Oracle версии 9i, корпорации Oracle (США). Преимуществами СУБД Oracle является то, что она работает под всеми распространенными в настоящее время операционными системами, не накладывает реальных ограничений ни на объемы хранимых данных, ни на количество одновременно работающих пользователей. Имеет мощные средства для создания хранимых процедур и средств поддержки целостности хранящихся данных. СУБД Oracle является объектно-реляционной СУБД, что позволяет эффектно использовать преимущества обеих парадигм при проектировании структуры БД и обработке данных в зависимости от специфики задачи.

Помимо всего прочего, в СУБД Oracle 9i имеется компонента Oracle Text, предназначенная для облегчения создания таких приложений, которые осуществляют поиск слов по текстам электронных документов. Для этого документы в БД индексируются (поддерживается инвертированный индекс по словам), при этом поддерживаются все распространенные форматы файлов (обычный текст, HTML, XML,

Microsoft Word и др.). Oracle Text содержит иерархическую БД категорий и понятий английского языка, тезаурус и список стоп-слов поиска.
Тип запроса пользователя определяет совокупность семантико-синтаксических шаблонов, которые будут рассматриваться на следующем шаге. Семантико-синтаксические шаблоны, используемые в системе, представляют собой формальные описания фраз или более крупных фрагментов текста, сконструированные с помощью описанных правил. В такие шаблоны могут включаться как конкретные слова, так и характеристики слов: смысл слова, грамматический класс или сорт. Совокупность этих шаблонов образует важную часть ЛБД.
Затем каждый шаблон преобразуется в набор предикатов SQL-запроса, для этого компоненты шаблона подвергаются предварительной обработке, заключающейся в раскрытии входящих в шаблон понятий и бинарных отношений, в формировании критериев отбора документов по наличию в них определенных слов и понятий. Преобразование осуществляется с помощью лексико-семантического словаря и словаря понятий Oracle Text. Полученный SQL-запрос выполняется для документов БД, при этом область его применения, в некоторых случаях, может быть сужена с помощью слов из запроса пользователя. Результат запроса подвергается дальнейшей обработке, зависящей от вида запроса пользователя.
При определении характера побочного эффекта от применения лекарственного препарата используется база знаний о медицине, т. к. изменение одного и того же фактора в одну и ту же сторону может носить как положительный, так и отрицательный характер, в зависимости от наличия определенных заболеваний у пациента.

Литература

1. Люстиг И. В. Основные принципы семантико-синтаксической обработки электронных документов в поисковой системе MEDSEARCH. // Научно-техническая конференция студентов, аспирантов и молодых специалистов МИЭМ. Тезисы докладов. - М.: МИЭМ. - 2004. - С. 292--294.
2. Fomichov V.A. A mathematical model for describing structured items of conceptual level // Informatica (Slovenia), Vol. 20, No. 1. P. 5-32.
3. Фомичев В.А. Математические основы представления смысла текстов для разработки лингвистических информационных технологий. Часть I. Модель системы первичных единиц концептуального уровня // Информационные технологии. 2002. № 10. С. 16-25.
4. Фомичев В.А. Математические основы представления смысла текстов для разработки лингвистических информационных технологий. Часть II. Система правил для построения семантических представлений фраз и сложных связных текстов // Информационные технологии. 2002. № 11. С. 34-45.