# РЕАЛИЗАЦИЯ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ КОНСТРУКЦИЙ В ИНТЕЛЛЕКТУАЛЬНОЙ ПОИСКОВОЙ СИСТЕМЕ MEDSEARCH

# AN IMPLEMETATION OF SEMANTIC-SYNTACTIC ANALYSIS OF NATURAL LANGUAGE CONSTRUCTIONS IN THE INTELLIGENT RETRIEVAL SYSTEM MEDSEARCH

***И. В. Люстиг***

***Московский государственный институт электроники и математики (технический университет), Москва***

***lyustig@mail.ru***

Поисковая система MEDSEARCH предназначена для получения справочной информации о лекарствах. Для ответа на некоторые виды запросов производится семантико-синтаксический анализ естественно-языковых конструкций. Для описания семантики лексических единиц использован аппарат теории стандартных К-языков.

Разработка поисковой системы MEDSEARCH ставит своей целью реализацию инструмента для интеллектуализации поиска справочной информации о лекарственных препаратах по базе данных их описаний. Система ориентирована на запросы следующих типов:

1) общие сведения о лекарственном препарате (ЛекП);
2) перечень ЛекП, применяемых для лечения данного заболевания;
3) список побочных эффектов от применения ЛекП;
4) характер возможных побочных эффектов с учетом особенностей пациента.

Проблема поиска информации в данном случае сводится к проблеме нахождения определенных сведений, сформулированных в описаниях ЛекП. Практика показывает, что поиск по ключевым словам, применяемый в большинстве поисковых машин, не дает высокого качества результата: число найденных документов велико, а их релевантность низка. Это обусловлено тем, что значительная часть слов в связных текстах на естественных языках (ЕЯ) лишь поддерживают связность и детальность описания и никоим образом не претендуют на выражение основного содержания документа или его фрагмента. Таким образом, формальное попадание ограниченного числа слов из запроса пользователя в текст документа не является гарантией его релевантности. Для обеспечения эффективного поиска знаний необходимо понимание поисковой системой как самого запроса, так и понимание ею электронного описания ЛекП в степени, достаточной для ответа на запрос пользователя [1]. Обеспечение возможности «понимания» запроса достигается с помощью заранее сформированных баз знаний о языке и предметной области, что позволяет строить по запросу пользователя поисковый образ релевантного документа или ЕЯ-конструкции, отвечающей на запрос.

Ядром системы является лингвистическая база данных (ЛБД), включающая лексико-семантический словарь (ЛСС) и совокупность семантико-синтаксических шаблонов, являющихся концептуальными моделями конструкций на ЕЯ. Методологической основой для построения этих шаблонов является теория К-языков, предложенная В. А. Фомичевым. Эта теория описывает математическую модель, задающую такие 10 операций на концептуальных структурах, с помощью которых можно строить семантические представления предложений и сколь угодно сложных связных текстов на ЕЯ [2–4], в том числе текстов по медицине [5]. При этом, т. к. многие объекты могут быть охарактеризованы с разных точек зрения — у них есть «координаты» по разным «семантическим осям», теория К-языков предоставляет возможность указания этих координат у семантических единиц, например, позволяет связать с конкретным пациентом «семантические координаты» «пространственный объект» и «интеллектуальная система».

Для ответа на запросы приведенных типов в системе MEDSEARCH оказалось достаточным использовать семантико-синтаксические модели смысловых отношений между словами предложения или нескольких предложений. Поиск смысловых отношений вместо использования семантико-синтаксических шаблонов фраз и более крупных фрагментов текста позволяет сократить объем ЛБД, уменьшить время обработки, расширить границы применимости метода поиска (обрабатываются предложения любой структуры, главное, чтобы они содержали интересующее смысловое отношение между определенными словами).

Используемые при поиске смысловые отношения описываются в ЛБД в словаре предложных и глагольно-предложных фреймов. Словарь предложных фреймов состоит из смысловых отношений вида «Существительное1 + Предлог + Существительное2», где предлог может быть пустым, и требований к их реализации.

В словаре глагольно-предложных фреймов описаны тематические роли, т. е. смысловые отношения между значением глагольной формы (личной формы глагола, неопределенной формы глагола, причастия, деепричастия, отглагольного существительного) и значением зависящей от нее в предложении группы слов. На формальном уровне тематические роли будут интерпретироваться, как названия бинарных отношений, где первым атрибутом является ситуация, а вторым — реальный или абстрактный объект, играющий определенную роль в этой ситуации.

Словарь глагольно-предложных фреймов в ЛБД содержит такие шаблоны (фреймы), которые позволяют представлять необходимые условия для реализации конкретной тематической роли в сочетании «Глагольная_форма + Предлог + Зависимая_группа_слов», где предлог может быть пустым, а зависимая группа слов является либо существительным с зависимыми словами или без них, либо конструктом, т. е. числовым значением параметра. Например, такими сочетаниями являются выражения «растворить в воде», «извлечь из упаковки», «принять лекарство», «содержать быстродействующее успокоительное», «заснуть до 22:00».

Работу поисковой системы MEDSEARCH с шаблонами смысловых отношений можно описать так. После получения на входе номера типа запроса и параметров запроса определяется совокупность смысловых отношений, которые будут искаться в текстах описаний ЛекП. Как уже упоминалось выше, смысловым отношениям в словарях ЛБД соответствуют шаблоны ЕЯ-конструкций, задающие требования к словам-участникам смыслового отношения для его реализации в тексте. Так, при поиске смысловое отношение, заданное в виде *tr = (item1; item2; requirement)*, считается присутствующим в тексте, если встречается пара семантических единиц *(item1, item2)* и выполняется требование *requirement* (содержит информацию о необходимом предлоге и грамматической форме связуемой семантической единицы). Затем для найденного смыслового отношения значение параметра поиска сравнивается с одной из семантических единиц, связанных отношением (выбирается исходя из имеющейся в системе информации, которая определяется в зависимости от типа запроса и от рассматриваемого смыслового отношения) на предмет того, что заданное значение параметра является конкретизацией данной семантической единицы.

При ответе на запрос первого типа (информация о ЛекП) система MEDSEARCH получает на входе название интересующего пользователя ЛекП, например, «Аспирин». Для ответа на запрос система формирует к БД SQL-запрос вида «Все документы, у которых в заголовке имеется слово «аспирин». Таким образом, запрос 1-го типа сводится к простому поиску по ключевому слову в некоторой части документа (в заголовке).

В запросе второго типа пользователь указывает название заболевания, которое должны лечить искомые ЛекП. Например, пользователь указал, что ищет все о лечении астмы. Для ответа на запрос система формирует к БД SQL-запрос вида «Все документы, у которых в разделе «Показания к применению» имеется слово «астма». Но это не все. Например, система могла найти предложение «Этот препарат не лечит астму, он предназначен для…», ведь в нем содержится слово «астма». Поэтому дальше необходимо проанализировать возвращенные запросом данные с помощью семантико-синтаксических шаблонов, а этот предварительный запрос выполнялся лишь для сокращения множества документов, которые необходимо проанализировать. Система должна отобрать те описания ЛекП, в которых в разделе «Показания по применению» реализуется смысловое отношение из набора смысловых отношений «Лечить заболевание» между описываемым препаратом и указанным пользователем заболеванием.

В качестве примеров ЕЯ-конструкций, где реализуются приведенные смысловые отношения, можно привести следующие фрагменты предложений из описаний ЛекП:

- «Симптоматическое лечение острых приступов бронхиальной астмы…»;
- «Применяется для лечения и ситуационной профилактики приступов удушья при астме…»;
- «Применяется при бронхиальной астме…».

При поиске понятия заменяются конкретными словами (с помощью ЛСС). Множество документов для лингвистического анализа сужается до описаний ЛекП, где близко (в одном предложении) встречаются оба слова с понятиями из рассматриваемого отношения (1-ое слово для одного понятия, 2-ое для второго понятия). Для каждой пары слов проверяется наличие между ними смыслового отношения.

Таким образом, для анализа ЕЯ-текстов в данном случае применяются шаблоны смысловых отношений, в которых фигурируют лишь ядерные понятия или слова (слова, наличие которые обязательно для реализации отношения и между ними нет слов-отрицаний).

Запрос третьего типа позволяет получать информацию о выявленных побочных эффектах от применения препарата. В данном случае пользователю необходимо указать тип запроса («Побочные эффекты») и ввести название

препарата. На основании введенной информации система выводит пользователю раздел «Выявленные побочные эффекты» из описания этого ЛекП, а также ищет вхождение названия этого препарата в раздел побочных эффектов в описаниях других лекарств. Поиск по другим описаниям обусловлен тем, что если пользователя интересует, например, аспирин, то его так же заинтересует ЛекП, в описании которого указано, что «Применение этого препарата одновременно с применением аспирина вызывает повышение кровяного давления», хотя в описании аспирина такой информации нет.

При анализе описаний других препаратов применяются смысловые отношения, выражающие одновременное применение препарата с рассматриваемым ЛекП. На настоящий момент это единственный тип искомых смысловых отношений по данному типу запроса, т. к. упоминание указанного лекарства в описаниях других препаратах интересно лишь в контексте их совместного применения.

Описания препаратов, в которых этот поиск дал положительный результат, подвергаются лингвистическому анализу для определения совместимости рассматриваемых лекарств при помощи смысловых отношений совместимости и несовместимости. Результаты анализа также сообщаются пользователю.

Четвертый тип запросов предназначен для получения информации о ЛекП, которые лечат данное заболевание, не усугубляя состояния пациента, страдающего рядом других заболеваний. Исходными данными для поиска являются название заболевания, которое необходимо вылечить, и перечень особенностей пациента, которые могут быть следующих типов:

1) название заболевания (например, сахарный диабет);
2) параметр с отклонением от нормы (например, повышенное кровяное давление);
3) непереносимость ЛекП или вещества (например, непереносимость новокаина).

При поиске сначала отбираются препараты, которые лечат указанное заболевание (по алгоритму обработки запроса 2-го типа). Затем для полученных препаратов проверяется их переносимость: названия данного лекарства нет в списке непереносимых пациентом препаратов и нет непереносимости к компонентам препарата (в состав препарата не входят непереносимые вещества), такая проверка осуществляется простым поиском по ключевым словам. ЛекП, для которых выявлена непереносимость, выводятся пользователю в виде дополнительной информации. Для оставшихся препаратов исследуется раздел «Побочные эффекты» и «Указания по применению» с целью выяснения, не усугубляет ли это лекарство имеющихся у пациента проблем со здоровьем, т. е. не вызывает отклонения указанных в особенностях пациента характеристик 2-го типа (например, «Кровяное давление — повышенное значение», «Пульс — пониженное значение») в ту же сторону (не усугубляет состояния пациента). При этом анализе используются смысловые отношения. Искомыми смысловыми отношениями будут: «Понижать — Симптом», «Повышать — Симптом», где вместо понятия «Симптом» будет подставляться заданная характеристика («кровяное давление», «пульс»), а тип искомого смыслового отношения будет зависеть от характера отклонения (если особенность больного «Кровяное давление — повышенное значение», то искомое отношение «Повышать — кровяное давление»), если отношение в тексте найдено, ЛекП считается непригодным, т. к. ухудшает состояние больного.

При ответе на запрос четвертого типа используется база знаний о медицине, которая позволяет определить, изменением каких факторов здоровья сопровождается каждое заболевание.

Реализация задачи семантико-синтаксического поиска для русского языка требует разработки морфологического анализатора. Морфологические словари достаточно большого объема (для русского языка количество лемм 162 519, количество наборов окончаний 2 553; в основе русского словаря лежит морфологический словарь Зализняка), пополняемые добровольцами, взяты с Интернет-сайта по адресу http://www.aot.ru, где они находятся в свободном доступе [6].

В качестве системы управления БД описаний ЛекП выбрана СУБД Oracle 9i. Эта СУБД имеет компонент Oracle Text, облегчающий создание приложений для поиска по текстам электронных документов: можно создавать инвертированный индекс по словам документов, поддерживаются все распространенные форматы файлов (обычный текст, HTML, XML, Microsoft Word и др.), имеется список стоп-слов поиска, существуют операторы для указания расстояния между искомыми словами запроса в тексте и проч.

*Пример или подпись к рисунку*

- Маркированный список.

4) Нумерованный список первого типа

а) Нумерованный список второго типа

### *Список литературы:*


1) Жигалов В. А. Составляющие интеллектуального поиска в сети // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог’2003 (Протвино, 11–16 июня 2003 г.) / Под ред. И. М. Кобозевой, Н. И. Лауфер, В. П. Селегея. М.: Наука, 2003. С. 749–756.
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. 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.
5) Люстиг И. В., Фомичев В. А. Принципы формального отображения семантики лексических единиц, предложений и дискурсов в интеллектуальной поисковой системе MEDSEARCH // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'2004 («Верхневолжский», 2–7 июня 2004 г.) / Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. М.: Наука, 2004. С. 431–435.
6) Сокирко А. В. Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'2004 («Верхневолжский», 2–7 июня 2004 г.) / Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. М.: Наука, 2004. С. 559–564.