УДК 681.3

ПРИМЕНЕНИЕ ОБЪЕКТНОЙ МОДЕЛИ ПРИ ПОИСКЕ ИЗВЛЕЧЕНИИ ИНФОРМАЦИИ ИЗ ЕЯ МЕДИЦИНСКИХ ТЕКСТОВ, ПРЕДСТАВЛЕННЫХ В ИНТЕРНЕТ

И.А. Коломойцева
Донецкий национальный технический университет

Стаття описує об'єктну модель природно-мовного медичного тексту з описом лікарських препаратів. Стаття містить приклади семантично значущих об'єктів і семантичних відносин, які можна зустріти в природно-мовному медичному тексті. Наведена схема програмної реалізації виділених об'єктів, понять і відносин.

Введение

В последнее время Интернет играет важную роль при получении информации в различных областях. А так как в Интернете хранится огромное количество разрозненной и во многих случаях повторяющейся информации, то её обработка требует автоматизации.

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 (медицинский ЕЯ-текст) к хорошо структурированной информации, которую можно обработать стандартными и высокоэффективными средствами информационных технологий.

В данной работе представлена схема алгоритма семантического анализа естественно-языковых текстов, содержащего описание лекарственных препаратов. Для этого определены объекты, которые присутствуют в медицинских ЕЯ-текстах с описаниями лекарственных препаратов и могут являться субъектами семантических отношений; определены семантические отношения (связи) для медицинского ЕЯ-текста. Также приведена схема программной реализации алгоритма.

1. Объекты и семантические отношения

Чтобы использовать естественный язык в качестве основы для построения языка представления знаний, в нем предлагается выделить несколько классов–элементов. Эти классы можно разделить на две категории: семантически значимые объекты предложения и

семантические отношения. Объекты еще называют именами [2] и именованными сущностями [3]. Примеры объектов, представленных в медицинских ЕЯ-текстах с описанием лекарственных препаратов, приведены в таблице 1.

Объекты связываются между собой с помощью семантических отношений. Выдвинута гипотеза, согласно которой множество отношений, в отличие от множеств объектов (имен), конечно [2]. В [2] выделено около 200 не сводимых к друг другу отношений. В [4] 200 отношений из [4] сведены к семнадцати. Более подробный обзор семантических отношений, определяемых для ЕЯ-текстов, представлен в [5, 6, 7].

Таблица 1. Объекты, представленные в медицинских ЕЯ-текстах

№ п/п	Название объекта	Примеры объектов
1	ЛЕКАРСТВО	Анальгин, аспирин, флемоксин
2	БОЛЕЗНЬ (ПОКАЗАНИЕ_К_ДЕЙСТВИЮ)	Аллергия, атеросклероз, остеохондроз, диабет сахарный
3	ПОБОЧНЫЕ_ДЕЙСТВИЯ	Тошнота, рвота, нарушение сна
4	ПРОТИВОПОКАЗАНИЯ	Аллергия, беременность
5	ИЗГОТОВИТЕЛЬ	Бристол-Майерс Сквибб
6	ФАРМГРУППА	Антибиотики, анальгетики
7	СОСТАВ	Бария сульфат

Классы-объекты можно представить в виде древовидной структуры (рисунок 1). На рисунке 1 представлен только небольшой фрагмент классов. Особенностью данного дерева является то, что в узлах дерева находятся названия классов, а листьями являются понятия данного класса, что позволяет достаточно четко их определять. Кроме того, все листья, которые определены в данном классе, являются синонимами.

В медицинских естественно-языковых текстах можно выделить следующие семантические связи: генеративную, результативную, инструментальную, каузальную, комитативную [5].

Генеративная связь имеет место, когда один компонент обозначает лицо или предмет, принадлежащий некоторой совокупности, категории, обозначаемой вторым компонентом.

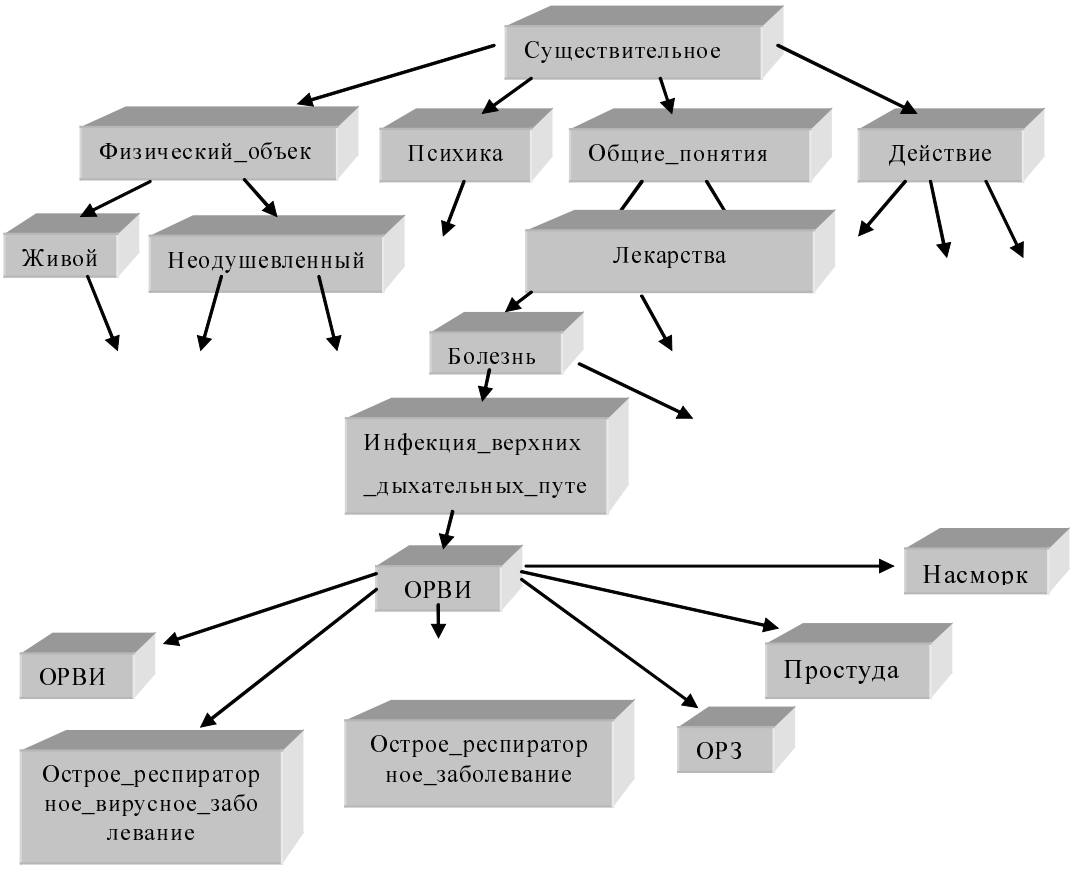

Рис. 1 Фрагмент классификации понятий в виде дерева

Таблица 2. Семантические отношения и связываемые ими объекты

Семантическая связь	Связываемые объекты
Результативная	ЛЕКАРСТВО → ПОБОЧНЫЕ_ДЕЙСВИЯ
Инструментальная	ЛЕКАРСТВО → БОЛЕЗНЬ
Каузальная	ЛЕКАРСТВО → ПОБОЧНЫЕ_ДЕЙСВИЯ
Комитативная	ЛЕКАРСТВО → ПРОТИВОПОКАЗАНИЯ

Результативная присутствует в тех предложениях, где один компонент выражает следствие действия второго.

Инструментальная означает, что один компонент обозначает орудие действия, обозначаемого другим компонентом.

Каузальная имеет место, когда один компонент обозначает причину появления другого компонента спустя какое-то время.

Комитативная встречается в тех предложениях, когда один компонент обозначает сопровождающее другой компонент действие, сопутствующий предмет, сопровождающее лицо.

Примеры объектов медицинских ЕЯ-текстов, которые связываются семантическими отношениями, представлены в таблице 2.

2. Основные принципы программной реализации семантического анализа ЕЯ-текста, содержащего описания лекарств, на основе объектно-понятийной модели текста

На вход системе подаётся информация следующего типа:

1) множество HTML-страниц, предположительно содержащих описания лекарств;

2) запросы пользователей в виде набора желаемых параметров (заболевания, список недопустимых противопоказаний);

3) запросы и команды администратора (эксперта) к базе данных.

На выходе система предоставляет:

1) множество найденных лекарств, соответствующих запросу пользователя

2) информация из базы данных, предоставляемая администратору (эксперту).

Система разбита на несколько модулей:

1) модуль поиска;

2) модуль анализа;

3) модуль модификации (обновления) базы знаний.

Модуль поиска получает из сети текст страницы и сохраняет ее в отведенном каталоге, одновременно обновляя данные в БД.

Параллельно модуль анализа анализирует уже полученные страницы, используя правила поиска из БД, и возвращает в базу найденные объекты.

Модуль модификации базы знаний изменяет схему классов, понятий и отношений при обнаружении новых объектов.

На уровне модуля поиска можно выделить два подмодуля: модуль поиска ссылок и модуль получения страниц. Модуль получения, используя данные из БД, закачивает HTML-страницы и сохраняет их в пуле, одновременно передавая на анализ модулю поиска ссылок, все найденные ссылки помещаются в БД и, в зависимости от настроек, обрабатываются.

База данных данного программного продукта представляется в виде множества таблиц, в которых находятся данные, используемые

для поиска страниц и анализа текста, а так же таблиц для хранения найденной информации.

Таблицы с данными для анализа можно представить в виде иерархичной структуры следующего вида: в таблицах нижнего уровня находятся ключевые слова и шаблоны - базовые лексемы, выделяемые из потока текста. Каждое слово определяется его корнем и количеством возможных символов в окончании, что позволяет расширить вариативность поиска без существенного увеличения затрат памяти и времени и при незначительном приросте погрешности. Шаблоны являются более сложными элементами текста и предназначены для поиска слабо нормированных лексем. Ключевые слова и шаблоны связаны с таблицей значений, которая содержит ограничения на значения параметров. Для тех параметров, которые принимают значения из ограниченного набора, в таблице храниться множество допустимых ключевых слов. Для других параметров указывается используемый шаблон поиска, ограничения на значения (минимум и максимум) и единица измерения. Таблица значений связана с таблицей параметров, в которой содержится информация об отдельных свойствах объекта поиска. Кроме множества ключевых слов, указывающих на параметр, в таблице содержится также номер группы значений, которые может принимать данный параметр, и данные, связанные с его положением в тексте и видом. Группы параметров объединяются в классы (например “действие”). Основная особенность классов – высокая вероятность близкого расположения друг с другом.

Вторая группа таблиц базы – таблицы, хранящие найденную информацию. Это таблица ссылок, содержащая список ссылок, у каждой ссылки имеется присоединенная строка – название скачанной страницы в пуле, и флаги, определяющие состояние ссылки. Таблица лекарств содержит информацию о найденных объектах.

Общая структура базы представлена на рисунке 2.

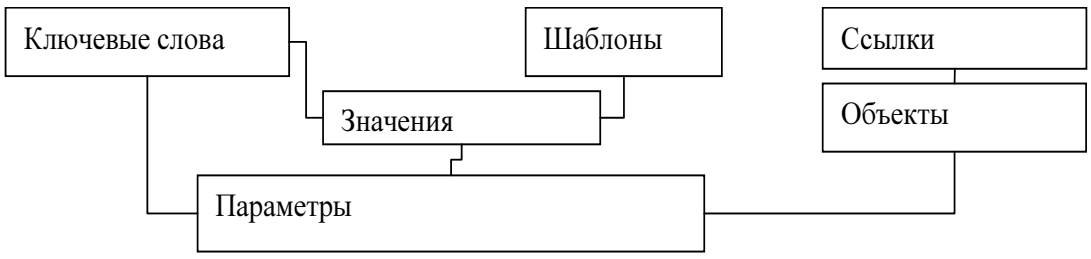

Рис.2 –Общая структура представления данных

Выводы

Выделенные в медицинском ЕЯ-тексте объекты и отношения могут служить в качестве словарей для организации семантического разбора. Объекты будут частью словаря перевода, а отношения – концептуального словаря.

Выделенные объекты и отношения можно использовать при создании программной системы автоматизированного поиска и извлечения знаний из медицинский естественно-языковых текстов, представленных в Интернет.

При дальнейшей работе с семантическими связями планируется определить их свойства.

Литература

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

3. 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.

4. Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. М.: Наука. Физматлит, 1997. 112 с.

5. Коломойцева И.А. Особенности применения существующих теорий «понимания» текста на естественном языке к медицинским текстам //Научные труды Донецкого государственного технического университета. Серия: Проблемы моделирования и автоматизации проектирования динамических систем, выпуск 29. Севастополь: «Вебер», 2001. С. 94–99.

6. Grishman. Information extraction: Techniques and challenges // Maria Teresa Pazienza, editor. Information Extraction. Springer-Verlag, Lecture Notes in Artificial Intelligence, Rome, 1997. P. 108-110.

7. Using a language independent domain model for multilingual information extraction. By: Azzam, Saliha; Humphreys, Kevin; Gaizauskas, Robert; Wilks, Yorick. Applied Artificial Intelligence, Oct 99, Vol. 13 Issue 7. P. 705-724.

Получено 27.05.09