УДК 004.89

КОМПЬЮТЕРНЫЙ ТЕЗАУРУС ДЛЯ ПРЕДМЕТНОЙ ОБЛАСТИ «ЛЕКАРСТВА»

И.А. Коломойцева
Донецкий национальный технический университет
kolomoit@pmi.dgtu.donetsk.ua

В данной работе представлено описание объектов и связей между объектами, которые выделены в предметной области «Лекарства». Также представлена структура и программная реализация тезауруса для рассматриваемой предметной области.

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 представлено описание объектов и связей между объектами, которые выделены в предметной области «Лекарства». Также представлена структура и программная реализация тезауруса для рассматриваемой предметной области.

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. Фрагмент этой древовидной структуры приведен в [5]. 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.

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

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.

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. Описание тезауруса предметной области

Знания о предметной области «Лекарства» необходимо представить в виде, пригодной для автоматической обработки. Онтологии являются именно такой формой представления знаний.

Онтология – соглашение об общем использовании понятий, которое содержит средства представления предметных знаний и договоренности о методах соображений. Она может рассматриваться как определенное описание взгляда на мир в конкретной сфере интересов, который состоит из набора терминов и правил использования этих терминов, которые ограничивают их значение в рамках конкретной предметной области [2].

Согласно «Современному словарю иностранных слов» тезаурус определяется как полный систематизированный набор данных о какой-либо области знаний, позволяющий человеку или вычислительной машине в ней ориентироваться. Таким образом, тезаурус можно рассматривать как частный случай онтологии.

Существует множество стандартов на формат представления тезаурусов. Основными документами, регламентирующими формат представления тезауруса, являются стандарты ISO 2788-1986 для описания одноязычных тезаурусов и ISO 5964-1985. При разработке тезауруса предметной области «Лекарства» будем опираться на первый из перечисленных стандартов.

Стандарт ISO 2788-1986 определяет тезаурус как набор терминов, связанных между собой соответствующими отношениями.

Каждый термин характеризуется комментарием и ссылкой на понятие верхнего уровня.

Основные понятия, представленные в тезаурусе «Лекарства»: название, лекарственная форма, фармакологическая группа, фирма-производитель, показания, противопоказания, заболевания, побочные эффекты, условия хранения, способ применения, действие.

Связи, которые представлены в тезаурусе «Лекарства»:

– USE – связь с очень близким по смыслу синонимом (термины A и B, представляющие некоторое понятие, обозначают практически одно и то же; например, у одного лекарства или одного заболевания может быть несколько названий);

– UF – связь, обратная USE;

– BT – связь термина с понятием высокого уровня (например, термин «гранулы» есть уточнение понятия «лекарственная форма);

– NT – связь, обратная BT;

– Res – Результативная;

– Ins – Инструментальная;

– Caus – Каузальная;

– Com – Комитативная.

Программно тезаурус представлен таблицами termin, comment и links.

Таблица termin содержит поля, в которых присутствует информация о названии понятия, ссылка на комментарий, ссылка на информацию о связях, идентификатор уровня.

Таблица comment содержит словесное представление комментария.

Таблица Links содержит информацию о типе связи, идентификаторы связываемых понятий.

Разработанный тезаурус можно использовать для программной системы автоматизированного поиска и извлечения знаний из медицинский естественно-языковых текстов, представленных в Интернет.

Выводы

Выделенные в медицинском ЕЯ-тексте объекты и отношения могут быть представлены в виде компьютерного тезауруса.

Разработанный тезаурус для предметной области лекартсва можно использовать для программной системы автоматизированного

поиска и извлечения знаний из медицинский естественно-языковых текстов, представленных в Интернет.

Список литературы

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.

8. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2000. – 384 с.

Получено 10.09.2011