УДК 681.3

Объектная модель естественно-языкового медицинского текста на примере системы «ФармАналитик»

Коломойцева И.А.
Донецкий национальный технический университет
kolomoit@r5.dgtu.donetsk.ua

Abstract

Kolomoitseva I. Object model of semantic analysis of natural language medical text on example of the «FarmAnalitik system». *An article describes an object model of natural language medical text, which contains description of medicinal preparations, place of this model in semantic interpretation of text. The article contains examples semantically meaningful objects and semantic relations which it is possible to meet in a natural-linguistic medical text. The example of the use of definite objects and relations in the «FarmAnalitik system» is resulted*

Введение

В последнее время все больше и больше специалистов 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.

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.

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.

Разработанный алгоритм реализован в системе «ФармАналитик» в подсистеме формирования базы данных лекарственных препаратов.

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]. В [3] также представлены объекты, которым оперирует процессор OntosMiner/Russian. 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. Остальные виды взаимосвязей между объектами, которые могут встретиться в естественно-языковом тексте, сводимы к этим базовым отношениям. В [4] 200 отношений из [2] сведены к семнадцати. В [2] и [4] тип связи определяется по тому, в каком падеже стоит объект, являющийся субъектом действия в предложении, и какой предлог предшествует этому объекту в предложении. Другой подход к определению семантических отношений предложен в [1]. В этом случае определено всего пять семантических отношений,

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

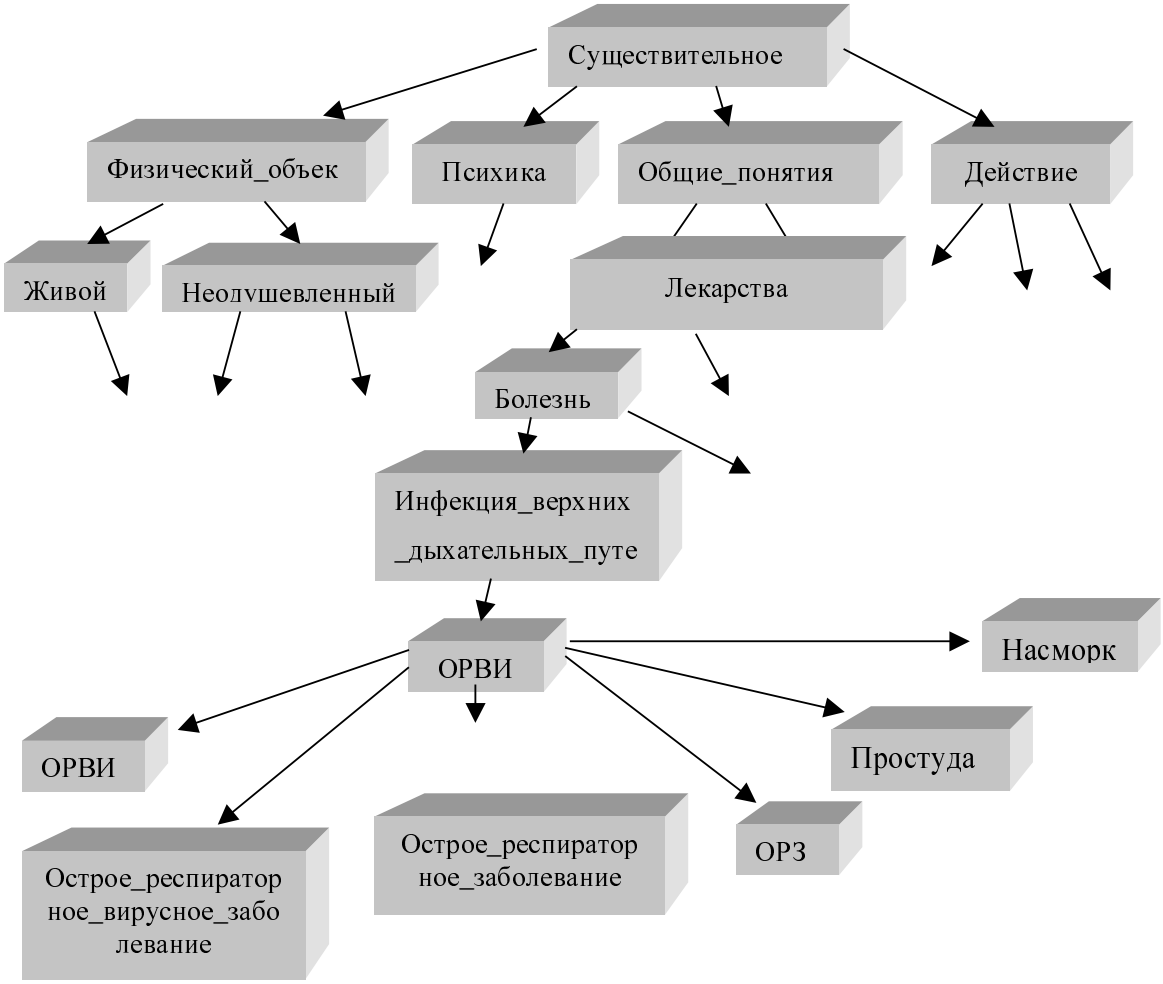

Рис. 1 Фрагмент классификации понятий в виде дерева

которые связывают между собой не объекты, а семантические классы. Семантические классы, в свою очередь, являются совокупностью объектов. Более подробный обзор семантических отношений, определяемых для ЕЯ-текстов, представлен в [5, 6, 7, 8, 9, 10].

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 [4].

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.

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

2. Система «ФармАналитик»

В рамках системы «ФармАналитик» выполняются следующие действия:

1) поиск в Интернет описаний лекарств (поиск происходит по списку сайтов, указанных пользователем);

2) извлечение информации из файлов с описанием препаратов и модификация базы данных лекарств и понятийных словарей (представленных в виде xml-файла);

3) поиск по базе данных лекарств, удовлетворяющих запросу пользователя.

Результатом работы поисковой подсистемы «ФармАналитика» является файл с описанием лекарства. Информация из этого файла подвергается анализу, результатом которого является запись БД, хранящая следующую информацию о лекарстве: название, список побочных действий, список противопоказаний и время действия. Если такого лекарства в базе данных нет, то полученная запись добавляется в общую базу данных. Если информация о найденном лекарстве есть, то после сравнительного анализа имеющейся информации и вновь полученной данные о лекарстве модифицируются.

Кроме этого добавляется информация в понятийные словари: обновляются классы-объекты «ЛЕКАРСТВА», «БОЛЕЗНИ», «ПОБОЧНЫЕ_ДЕЙСВИЯ» и «ПРОТИВОПОКАЗАНИЯ».

При поиске лекарства пользователь системы «ФармАналитик» вводит название болезни и список недопустимых противопоказаний. Поиск проходит не только по указанному заболеванию, но и по всему синонимическому ряду. Результатом данного запроса будет список отобранных лекарств

Следующим шагом является сравнительный анализ лекарств.

На вход системе поступает два параметра: максимальное количество сравниваемых лекарств и сам список выбранных лекарств для сравнения.

Количество сравниваемых лекарств может варьировать от 2 до 5. Причем пользователь сам может задавать этот показатель.

Порог слева равен двум, иначе сравнение не будет иметь смысла. Максимальное количество сравниваемых лекарств достигает пяти.

При выводе списка лекарств применяется сортировка по возрастанию по обобщённому показателю, учитывающему количество противопоказаний и побочных действий, поэтому самое первое лекарство будет наилучшим. Заканчивается работа программы выводом на экран информации о лекарствах и их характеристиках.

Таким образом, пользователь получает название и параметры наилучшего лекарства, отобранного системой, и параметры всех остальных сравниваемых лекарств.

Заключение

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.

При дальнейшей работе с семантическими связями планируется определить их свойства.

Так как разумно построенная система анализа должна обеспечивать не только извлечение знаний из конкретного текста, но и накопление результатов, как на синтаксическом, так и на семантическом уровне [1], то в качестве перспектив дальнейшей работы можно указать следующее:

1) создание большого корпуса медицинских ЕЯ-текстов с целью их анализа и уточнения количественного и качественного состава семантически нагруженных объектов;

2) определение необходимых и достаточных условий для установления типа семантической связи;

3) модификация алгоритма автоматического выделения классов-объектов из естественно-языкового медицинского текста.

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.
8. Nirenburg S., Raskin V. Ontological Semantics. – Cambridge, MA: MIT Press, 2004 – 350 p.
9. Rosse C, Ben Said M, Eno KR, Brinkley JF. Enhancements of anatomical information in UMLS knowledge sources. Proc Annu Symp Comput Appl Med Care. 1995; p. 873-887.
10. Halper MH, Chen Z, Geller J, Perl Y. A metaschema of the UMLS based on a partition of its semantic network. Proc AMIA Symp. 2001;: p. 234-238.

Поступила в редколлегию 13.03.2009