Объектная модель семантического анализа естественно-языкового медицинского текста

Коломойцева И.А.

Донецкий национальный технический университет
kolomoit@pmi.dgtu.donetsk.ua

Abstract

Kolomoitseva I.A. Object model of semantic analysis of natural language medical text. An article describes an object model of natural language medical text, place of this model in the semantic analysis of text. An article contains the examples semantically meaningful objects and semantic relations, which it is possible to meet in the natural language medical text. In article is indicated, that semantic objects and relations are a basis of dictionaries, which participate in the semantic analysis of text.

Введение

В последнее время все больше и больше срециалистов 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].

Среди теоретических проблем, которые существуют в компьютерной семантике, наименее разработанными являются следующие проблемы [1]:

1) стандартизация языков представления знаний;

2) разрешение синтаксической и лексической омонимии;

3) установление референциальных отношений между единицами текста;

4) анализ контекстов, характеризующихся смысловой неполнотой;

5) разработка семантических словарей, необходимых для поддержки алгоритмов семантического анализа;

6) реализация логического вывода по тексту.

В целом стоит отметить, что компьютерная семантика только выходит из стадии поисковых и научно-исследовательских работ.

Области применения семантического анализа очень разнообразны. Например [1]:

1) формирование дополнительных лингвистических фильтров в системах распознавания (OCR и Speech Recognition) и в грамматических корректорах;

2) разрешение неоднозначностей и повышение уровня профессиональной компетенции в системах перевода;

3) формирование дополнительных критериев релевантности документа в документальных ИПС;

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

Методы и средства семантического анализа можно разделить на два направления [1]:

1) средства формализации фактологической информации (для СУБД);

2) средства формализации номологической информации (для экспертных систем).

Именно 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.

Семантический анализ ЕЯ-текста опирается на концептуальный и понятийные словари [1]. Понятийный словарь включает в себя описание понятий (объектов), которые представлены в тексте. Назначение этого словаря – установить связь между понятиями (объектами) и конкретными словами, встреченными в тексте. Концептуальный словарь оперирует смыслами, описывает свойства и отношения понятий, то есть семантические связи. Стратегия наполнения этих словарей является актуальной научной задачей.

Целью работы является:

1) определение объектов, которые присутствуют в медицинских ЕЯ-текстах, могут являться субъектами семантических отношений и будут использоваться при создании семантических словарей;

2) определение семантических отношений (связей) для медицинского ЕЯ-текста с использованием выявленных объектов;

3) определение роли семантически значимых объектов и семантических отношений в структуре словарей, используемых для семантического анализа медицинского ЕЯ-текста.

Постановка задачи

Задачей данной работы являвется анализ текстов, полученных в результате поиска в сети Интернет. Эти тексты имеют следующую структуру:

– название болезни;
– общее описание болезни;
– симптомы и причины возникновения;
– методы лечения.

Результаты анализа медицинского ЕЯ-текста должны быть представлены в виде семантически значимых объектов и семантических отношений. Выявленные объекты и отношения станут основой понятийного и концептуального семантических словарей.

Общие положения семантического анализа

Ограничения, накладываемые на текст, по которому проводится семантический анализ [1]:

1) в тексте должен присутствовать смысл и этот смысл должен быть ярко выражен;

2) тексты должны опираться на логически и терминологически отработанную систему понятий;

3) тексты должны быть стилистически и лексически однородными;

4) тексты должны быть фактографическими, содержать описание свойств определенной совокупности объектов, отношений между ними, процессов и действий, в которых они участвуют.

Выход за пределы этих ограничений требует применения специальных методов, ориентированных на специфику решаемой задачи.

На вход семантического компонента должен поступать синтаксически размеченный текст (унификация разметки – задача полностью не решенная) [1]. В размеченном тексте должна быть представлена следующая информация:

- идентификаторы понятий, соответствующих слову (термину);
- указание синтаксического хозяина и вида синтаксической связи;
- выделение сегментов (части сложного предложения, обособленных оборотов);
- раздельное представление всех глобальных вариантов синтаксического разбора;
- анафорические ссылки;
- дополнительная грамматическая информация о слове, которая может потребоваться в процедурах семантического анализа.

Также должны быть опознаны и представлены одной лексемой термины-словосочетания, унифицировано представление числовой информации, опознаны собственные имена и т.п.

Система семантического анализа является тезаурусно (онтологически) ориентированной. Основная проблема в создании реально работающих анализаторов - это создание реально работающего понятийного словаря, то есть словаря, который:

1) обеспечивает требуемую алгоритмами функциональность;

2) обеспечивает удовлетворительное покрытие текстов.

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. Объекты, представленные в медицинских ЕЯ-текстах, описаны в таблице 1. Так как объект ЭЛЕМЕНТЫ из таблицы 1 подразумевает довольно много понятий, то в нем можно выделить ряд подобъектов, которые представлены в таблице 2.

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, 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] .

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.

Таблица 1. Объекты, представленные в медицинских ЕЯ-текстах

№ п/п	Название объекта	Примеры объектов
1	БОЛЕЗНЬ	Аллергия, атеросклероз, остеохондроз, диабет сахарный
2	КАЧЕСТВА_ЛЮДЕЙ	Возраст, профессия, занятия
3	НАРУШЕНИЯ	Нарушения работы организма человека: авитаминозы
4	ЭЛЕМЕНТЫ	Вещества, которые участвуют в жизнедеятельности человеческого организма
5	СОСТОЯНИЕ_ОРГАНИЗМА	Белковое голодание, витаминная недостаточность, гиподинамия, гиповитаминоз, улучшение состояния организма (лечение), азотное равновесие
6	НАУКА	Биохимия
7	СИСТЕМА_ПИТАНИЯ	Вегетарианство, диета, голодание
8	ДЕЙСТВИЕ_НА_ОРГАНИЗМ	Липотропное воздействие
9	ОБРАЗ_ЖИЗНИ	Сидячий, активный
10	МЕТОД_ЛЕЧЕНИЯ	Лекарство, операция, мануальная терапия
11	СИМПТОМЫ	Боли в органах

Таблица 2. Подобъекты объекта ЭЛЕМЕНТЫ

№ п/п	Название подобъекта	Примеры подобъектов
1	ВИТАМИНЫ	Водорастворимые, жирорастворимые, холин
2	ГОРМОНЫ	Инсулин
3	БЕЛКИ	
4	ЖИРЫ	Гидрированные, липиды
5	УГЛЕВОДЫ	Гликоген, глюкоза, дисахариды, моносахариды, полисахариды
6	ФЕРМЕНТЫ	
7	КИСЛОТЫ	Жирные кислоты (насыщенные, ненасыщенные, полинасыщенные), аминокислоты (метионин, цистин)
8	ВЕЩЕСТВА	Фитонциды, балластные вещества (клетчатка, пектин), холестерин)

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.

Объекты медицинских ЕЯ-текстов, которые связываются семантическими отношениями, представлены в таблице 3.

Таблица 3. Семантические отношения и связываемые ими объекты

Семантическая связь	Связываемые объекты
Результативная	ОБРАЗ_ЖИЗНИ → БОЛЕЗНЬ ОБРАЗ_ЖИЗНИ → СОСТОЯНИЕ_ОРГАНИЗМА СИСТЕМА_ПИТАНИЯ → БОЛЕЗНЬ БОЛЕЗНЬ → СОСТОЯНИЕ_ОРГАНИЗМА БОЛЕЗНЬ → БОЛЕЗНЬ МЕТОД_ЛЕЧЕНИЯ → СОСТОЯНИЕ_ОРГАНИЗМА НАРУШЕНИЯ → СОСТОЯНИЕ_ОРГАНИЗМА НАРУШЕНИЯ → НАРУШЕНИЯ
Инструментальная	НАУКА → МЕТОД_ЛЕЧЕНИЯ НАУКА → СИСТЕМА_ПИТАНИЯ МЕТОД_ЛЕЧЕНИЯ → НАРУШЕНИЯ
Каузальная	ЭЛЕМЕНТЫ → СОСТОЯНИЕ_ОРГАНИЗМА ЭЛЕМЕНТЫ → БОЛЕЗНЬ СИСТЕМА_ПИТАНИЯ → ЭЛЕМЕНТЫ ОБРАЗ_ЖИЗНИ → БОЛЕЗНЬ КАЧЕСТВА_ЛЮДЕЙ → БОЛЕЗНЬ КАЧЕСТВА_ЛЮДЕЙ → СОСТОЯНИЕ_ОРГАНИЗМА ОБРАЗ_ЖИЗНИ → СОСТОЯНИЕ_ОРГАНИЗМА МЕТОД_ЛЕЧЕНИЯ → МЕТОД_ЛЕЧЕНИЯ
Комитативная	БОЛЕЗНЬ → СИМПТОМЫ БОЛЕЗНЬ → НАРУШЕНИЯ НАРУШЕНИЯ → НАРУШЕНИЯ

Пример анализа медицинского ЕЯ-текста

В качестве примера взято описание заболевания остеохондроз позвоночника, представленное на рисунке 1, в нем выделены семантически нагруженные объекты и определены связывающие их семантические отношения.

Остеохондроз позвоночника
Остеохондроз позвоночника это заболевание связанное с дегенерацией межпозвоночных дисков. Дегенерация дисков приводит к потере ими необходимой эластичности и в последствии к образованию межпозвоночных грыж, которые в свою очередь могут привести к защемлению нерва и соответственно болям, как в спине, так и в отдельных органах, за работу которого отвечает защемленный нерв.
Судя по некоторым медицинским публикациям, остеохондрозом позвоночника страдает до 80% населения, однако не во всех случаях остеохондроз приводит к такому осложнению как грыжа межпозвоночного диска и соответственно постоянным болям в спине. Однако число жалоб на боли в спине постоянно возрастает, и что не мало важно, все чаще остеохондроз позвоночника встречается у молодых людей. И хотя точные причины развития остеохондроза определить очень сложно, так как он встречается у людей разного возраста, профессий и образа жизни, то некоторые тенденции все же можно отметить. Зачастую, остеохондроз позвоночника встречается у людей проводящих долгое время в сидячем положении. Это в особенности относиться к водителям и пользователям персональных компьютеров. А связанно это с тем, что человек находиться впрактически неподвижном положении достаточно долгое количество времени.
Таким образом, можно сказать, что в большинстве случаев для профилактики и лечения остеохондроза эффективным будет ведение достаточно активного образа жизни, включающего в себя лечебную гимнастику, ходьбу, плавание. Так же, для лечения остеохондроза часто применяется мануальная терапия (массаж). По средствам массажа вы так же можете придать тонус мышцам спины, но надо иметь ввиду, что мануальная терапия может вызвать серьезные осложнения при наличии межпозвоночных грыж и в случае непрофессионализма мануального терапевта (массажиста). Более радикальным способом лечения остеохондроза позвоночника является операция. Операцию, как правило назначают при наличии межпозвоночной грыжи размером более 5 мм. и в случаях серьезного нарушения работы других органов вызванного защемлением нерва. Так же, операция назначается в случае отсутствия положительного результата в течении нескольких месяцев при применении консервативного лечения остеохондроза.

Рисунок 1 – Пример естественно-языкового медицинского текста

Результат проведенного анализа представлен в таблице 4.

Таблица 4 Результат анализа ЕЯ-текста, содержащего описание заболевания остеохондроз позвоночника

Семантическая связь	Связываемые объекты	Примеры связываемых объектов из ЕЯ-текста
Комитативная	БОЛЕЗНЬ → НАРУШЕНИЯ	Остеохондроз → Дегенерация межпозвоночных дисков
Результативная	НАРУШЕНИЯ → НАРУШЕНИЯ	Дегенерация межпозвоночных дисков → Потеря эластичности межпозвоночных дисков
Результативная	НАРУШЕНИЯ → НАРУШЕНИЯ	Потеря эластичности межпозвоночных дисков → Грыжа
Результативная	НАРУШЕНИЯ → НАРУШЕНИЯ	Грыжа → Защемление нерва
Результативная	НАРУШЕНИЯ → СОСТОЯНИЕ_ОРГАНИЗМА	Защемление нерва → Боли в спине и других органах
Каузальная	ОБРАЗ_ЖИЗНИ → БОЛЕЗНЬ	Проводить долгое время в сидячем положении → Остеохондроз
Каузальная	КАЧЕСТВА_ЛЮДЕЙ → БОЛЕЗНЬ	Водители → Остеохондроз
Каузальная	КАЧЕСТВА_ЛЮДЕЙ → БОЛЕЗНЬ	Пользователи ПК → Остеохондроз
Каузальная	ОБРАЗ_ЖИЗНИ → СОСТОЯНИЕ_ОРГАНИЗМА	Активный образ жизни → Улучшение состояния организма (профилактика и лечение)
Каузальная	ОБРАЗ_ЖИЗНИ → СОСТОЯНИЕ_ОРГАНИЗМА	Ходьба → Улучшение состояния организма (профилактика и лечение)
Каузальная	ОБРАЗ_ЖИЗНИ → СОСТОЯНИЕ_ОРГАНИЗМА	Плавание → Улучшение состояния организма (профилактика и лечение)

Продолжение таблицы 4.

Семантическая связь	Связываемые объекты	Примеры связываемых объектов из ЕЯ-текста
Каузальная	ОБРАЗ_ЖИЗНИ → СОСТОЯНИЕ_ОРГАНИЗМА	Мануальная терапия (массаж) → Улучшение состояния организма (профилактика и лечение)
Каузальная	ОБРАЗ_ЖИЗНИ → СОСТОЯНИЕ_ОРГАНИЗМА	Мануальная терапия (массаж) → Повышение тонуса мышц
Каузальная	КАЧЕСТВА_ЛЮДЕЙ → СОСТОЯНИЕ_ОРГАНИЗМА	Непрофессионализм мануального терапевта (массажиста) → Осложнения при наличии межпозвоночной грыжи
Инструментальная	МЕТОД_ЛЕЧЕНИЯ → НАРУШЕНИЯ	Операция → Межпозвоночная грыжа более 5 мм
Инструментальная	МЕТОД_ЛЕЧЕНИЯ → НАРУШЕНИЯ	Операция → Нарушения работы органов организма, вызванные защемлением нерва
Каузальная	МЕТОД_ЛЕЧЕНИЯ → МЕТОД_ЛЕЧЕНИЯ	Отсутствие положительного результата при применении консервативного лечения → Операция

Выводы

Анализ произвольно взятого медицинского текста показывает, что методика представления текста, предложенная в статье, применима к медицинским ЕЯ-текстам.

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.

Отличие предложенного подхода к представлению медицинского ЕЯ-текста от онтологического состоит в том, что в тексте выделяются не только понятия (объекты), но и выполнена попытка установить связь

между объектами в виде семантических связей. 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) разработка системы пополнения правил семантического разбора.

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.