ОСОБЕННОСТИ ПРИМЕНЕНИЯ СУЩЕСТВУЮЩИХ ТЕОРИЙ «ПОНИМАНИЯ» ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ К МЕДИЦИНСКИМ ТЕКСТАМ

Коломойцева И. А.
Кафедра ПМиИ, ДонГТУ
kolomoit@r5.dgtu.donetsk.ua

Abstract

TextAnalysis Theories Application Peculiarities On Natural Language To Medical Text. Kolomoitseva I. A *This article is about analysis of being texts understanding methods on natural Russian. In detail analysis approaches to this problem by D. A. Pospelov and G. S. Osipov. A Natural text examines as formal deductive system. Lead semantic intercourses, which one can be picked out in text on natural language. Comes to the conclusion of application of these methods to «understanding» of medical text on natural language.*

Введение

Процесс автоматизации принятия решения в экспертных системах невозможен без привлечения информации, которая не может быть выражена количественно. Это семантическая (смысловая) информация. Такую информацию возможно извлечь из естественноязыковых текстов, в случае медицинских экспертных систем знания извлекаются из специальных медицинских текстов, зафиксированных рассказов врачей-экспертов о различных проявлениях заболеваний, методов их лечения и т. п.

На протяжении последних двух-трех десятилетий многие исследователи, занимающиеся проблемами автоматической обработки текста и «понимания» естественного языка, получили ряд интересных результатов. Особенно продвинулись исследователи, занимающиеся «пониманием» текстов на английском языке. Это обуславливается довольно строгой структурированностью этого языка, выражающейся, в частности, например, закрепленным положением членов предложения (таких как подлежащее, сказуемое, обстоятельства, дополнения и т. п.). Классическими в этой области являются работы Роджера Шенка. Еще в 70-е годы он предложил стройную схему «понимания» текстов на английском языке [1]. Схема эта с годами улучшается, например, французские исследователи разработали свой подход к «пониманию» опять же текстов на английском языке, в основу которого была положена логика предикатов [2].

«Понимание» текстов на русском языке затрудняется сложной организацией этого языка. Однако достигнуты успехи и в этом направлении. Это заслуга многих советских фоссийских) ученых, среди которых можно выделить Д. А. Поспелова, Г. С. Осипова [3, 4]. В этой статье предпринята попытка применить их методики к медицинскому тексту на естественном языке.

Начнем с того, что любой текст на естественном языке можно представить как формальную дедуктивную систему.

1. Определение процесса формирования базы знаний медицинской экспертной системы как формальной дедуктивной системы

Формальная дедуктивная система может быть представлена как совокупность четырех множеств [3]

$$Ф=<Т, Р, А, П>, \quad (1)$$

где Т - базовые элементы - элементы, которые не могут быть поделены на более мелкие элементы (подэлементы или субэлементы);

Р - синтаксические правила, которые определяют, как из множества Т образуются производные элементы (Р является одной или множеством эффективных процедур, с помощью которых определяется правильность построения производных элементов относительно любой совокупности базовых; те элементы, которые строятся из Т с помощью Р, называются правильно построенными совокупностями или сокращенно ППС);

А - аксиомы, являющиеся некоторым произвольным подмножеством множества производных элементов формальной системы, которое построено в соответствии С Р (аксиомы ничем не отличаются от других ППС);

П - правила вывода, с помощью которых из аксиом можно вывести другие ППС, ранее не являвшиеся аксиомами (общая форма записи правил П имеет вид: K|=>Q;K - это совокупность ППС либо входящих в А, либо выведенных из А; Q - совокупность ППС, которые выводятся на данном шаге вывода).

Формальная система приводится к семиотической путем процедуры интерпретации, действие которой выражается в том, что множествам Т, Р, А и П придается реальный смысл относительно того объекта или процесса, который формализуется.

Для естественноязыкового медицинского текста базовыми элементами (множеством Т) являются буквы русского, латинского и греческого алфавитов, десятичные цифры и знаки: «.», «,», «!», «:», «?», «;», «-», «(», «)», «%», «»», ««», « »). Множество синтаксических правил Р имеет следующий вид: 1) буква русского алфавита есть ППС; 2) буква латинского алфавита есть ППС; 3) буква греческого алфавита есть ППС; 4) цифра есть ППС; 5) знак препинания есть ППС; 6) если Ω - ППС, а ω - последний символ из Ω, то к ω справа можно присоединить букву русского, латинского или греческого алфавитов, цифру или знак препинания; 7) других ППС нет.

Эти правила можно также описать в нотации Бэкуса-Наура.

<БУКВА>::=А|Б|В|Г|… | Э | Ю | Я | а | б | в | г | … | э | ю | я | A | B | C | D | … | X I Y I Z I a I b I c I d | … | x | y | z | A | B | X | Δ | … |Ψ | ς | Ω | α | β | χ | δ [ψ | w | ω

<ЦИФРА>::=0 [1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

<ЗНАКИ>::=,|.|-|;|:|!|?|(|)|<|>|{|}|[|]|«|»|ПРОБЕЛ|%

<СЛОВО>::=<БУКВА>|<СЛОВО><БУКВА>

<ЧИСЛО>::=<ЦИФРА>|<ЧИСЛО>< ЦИФРА >

<СЛОВ01>::=<СЛОВО>|<ЗНАК><СЛОВО>|<СЛОВОIхЗНАК>|<СЛОВО><ЗНАК>

<ЧИСЛ01>::=<ЧИСЛ0>|<ЗНАК><ЧИСЛ0>|<ЧИСЛ0><ЗНАК>|<ЧИСЛ01><ЗНАК>

<ПРЕДЛОЖЕНИЕ>::=<СЛОВО 1>|<ПРЕДЛОЖЕНИЕ><СЛОВО 1>|<ЧИСЛО 1>|<ПРЕДЛОЖЕНИЕ><ЧИСЛО 1>

<ТЕКСТ>::=<ПРЕДЛОЖЕНИЕ>|<ТЕКСТ><ПРЕДЛОЖЕНИЕ>

Применительно к медицинекому тексту формальным системам можно приписать следующие свойства:

1) с помощью правил и аксиом можно создать как генератор, так и разборщик медицинского текста;

2) каждая следующая аксиома генерируется независимо от порядка получения уже имеющихся аксиом;

3) множества аксиом и правил остаются неизменными в процессе вывода;

4) синтаксические правила оказывают влияние на выбор аксиом, так как аксиомы являются ППС;

5) интерпретация формальной системы остается неизменной при фиксированном множестве правил интерпретации, присваивающем смысл всем ППС в рамках медицинской предметной области.

Для семиотических систем не выполняются свойства 1)-3) формальных систем, так как интерпретатор при формировании базы знаний может изменять правила вывода и систему аксиом.

Причисление системы, где присутствует база знаний и интерпретатор, к семиотическим (знаковым) системам, правомерно, так как ее можно представить в виде [3]:

$$M=<Z,R> \quad (2)$$

где Z - множество знаков,

R - множество отношений между ними.

Под знаком в этом случае понимаются элементы, обладающие одновременно тремя свойствами: синтаксисом, семантикой и прагматикой. Отношения между этими свойствами неоднозначны.

Естественный язык представляет собой сложно организованную систему, которую можно рассматривать как семиотическую, так как в случае естественноязыкового текста мы имеем дело со знаками (буквами, цифрами, знаками препинания), которые располагаются в тексте в соответствии с определенными синтаксическими правилами (эти правила и образуют множество R, то есть отношения между символами языка). Кроме этого, знакам, образующим естественноязыковый текст, люди приписывают некий смысл; таким образом, эти знаки обладают свойством семантики.

Синтаксический уровень естественного языка давно формализован. Подробное описание такого формализмадля английского языка приведено в работах Р. Шенка [1], а для русского - в работах Д. А. Поспелова и Г.С. Осипова [3], [4]. Однако следуя лишь синтаксическим правилам русского (как и любого естественного) языка, всегда можно породить бессмысленную фразу. Продемонстрируем это на примере. Возьмем три семантически и синтаксически верных высказывания. 1) «Больной лечится в больнице». 2) «Лекарство готовится в аптеке». 3) «Осмотр проводится в школе». Все эти фразы имеют структуру:

N-G-E-M,

где N - множество существительных единственного числа в именительном падеже;

G - множество глаголов несовершенного вида, единственного числа, третьего лица, настоящего времени, стоящих в изъявительном наклонении;

E - множество непроизводных предлогов, относящихся к предложному падежу;

M - множество существительных единственного числа в предложном падеже.

Зададим множество T=T1uT2uT3uT4. Подмножества Tj содержат слова русского языка, относящиеся к множествам N, G, E, M соответственно. Если порождать ППС в этой формальной системе, то при наличии в Ti, **T2, T3,** T4 необходимых слов русского языка можно получить и приведенные выше фразы. Но такая система может породить, например, и следующие фразы: «осмотр лечится в аптеке», «больной готовится в больнице», «лекарство лечится в школе». Очевидно, что эти синтаксически правильно построенные фразы лишены смысла.

Из приведенного выше можно сделать вывод, что главное в формализации естественного языка - это формализация его семантической составляющей. Этот процесс значительно осложняется омонимией (многозначностью) элементов естественного языка. Именно этим объясняется то, что до настоящего времени ни один из естественных языков (а особенно русский) не удалось с необходимой полнотой описать не семантическом уровне [3].

В случае конкретной предметной области, в данном случае медицины, мы имеем дело с текстами, которые составлены не только в соответствии с синтаксическими, но и с определенными семантическими правилами, то есть по шаблонам. Это в какой-то мере облегчает их интерпретацию. Например, шаблонной является фраза: «При заболевании ... встречаются симптомы: ...», «Больной, страдающий ... (указывается заболевание), может испытывать боли в ...» и т. п.

2. Функциональные классы в естественныхязыках

Чтобы использовать естественный язык в качестве основы для построения языка представления знаний, в [3] предлагается выделить в нем несколько классов-элементов (слов и сочетаний), которые играют «определенную функциональную роль в представлении знаний». Наиболее важными из этих классов Д. А. Поспелов в [3] называет понятия, имена и отношения. В свою очередь «понятия» разбиваются на «понятия-классы», «понятия-процессы» и «понятия^остояния». Применим эту методику к медицинскому тексту. Понятие-класс определяется как совокупность объектов, обладающих вполне определенными свойствами, например, «лекарство», «больница», «врач». Понятия-процессы описывают группу однородных процессов. Например, «лечение», «приготовление лекарств», «постановка диагноза». Понятия-состояния близки по смыслу к понятиям-процессам и описывают какое-то определенное состояние. Например, «человек болен», «человек здоров», «человек находится в стадии выздоровления». И понятия-классы, и понятия-процессы, и понятия^остояния имеют имена, которые уточняют элементы, входящие В ТО или иное понятие. Например, «больница №2», «врач Иванов». В первом случае «именем» является номер больницы, во втором - фамилия врача. Очевидно, что множества понятий и имен для естественных языков бесконечны.

Отношения связывают между собой элементы множества понятий или идентифицированных понятий. Выдвинута гипотеза, согласно которой множество отношений, в отличие от множеств понятий и имен, конечно. Предполагается наличие около 200 не сводимых к друг другу отношений. Остальные виды взаимосвязей между

элементами множества понятий, которые могут встретиться в естественноязыковом тексте, сводимы к этим базовьм отношениям. Из множества базовых отношений, определенных в [3], можно выделить элементы, наиболее широко представленные В медицинских текстах. Результаттакого отбора представлен в таблице 1. Эти отношения выражают взаимосвязь объектов реального мира, а не элементов языка.

Таблица 1 Базовые отношения, часто встречающиеся в медицинских тестах

Тип отношения	Наименования отношения
Временные	Быть одновременно
	Быть раньше
Классификационные	Быть элементом класса
	Обладать
Идентифицирующие	Иметь имя
Прагматические	Служить для
	Обладать состоянием
	Участвовать в процессе
Каузативные	Быть целью
	Быть причиной
	Действие^збъект

Таблица 2 Семантические связи

Наименование семантической связи	Обозначение семантической Связи	Описание семантической связи
Генеративная	Gen	Один компонент обозначает лицо или предмет, принадлежащий некоторой совокупности, категории, обозначаемой вторым компонентом
Дестинативная	I)es	Один компонент обозначает назначение для другого компонента
Директивная	^ k	Один компонент обозначает путь, направление второго компонента
Инструментальная	Ins	Один компонент обозначает орудие действия, обозначаемого другим компонентом
Каузальная	Cous	Один компонент обозначает причину появления другого компонента спустя какое-то время
Комитативная	Com	Один компонент обозначает сопровождающее другой компонент действие, сопутствующий предмет, сопровождающее лицо
Коррелятивная	~Cor	Один компонент выражает возможность наблюдения другого компонента или соответствия предмета другому предмету, назначению
Негативная	Neg	Один компонент отрицает, исключает возможность проявления другого компонента

Продолжение таблицы 2

Лимитативная	Lim	Один компонент обозначает сферу применения, назначения другого компонента
Медиативная	Med	Один компонент имеет значение способа, средства действия другого
Поссесивная	Tos	Один компонент выражает отношение владения другим компонентом
Потенсивная	Tot	Один компонент приводит к увеличению возможности появления другого спустя некоторое время
Результативная	Tes	Один компонент выражает следствие действия второго
Репродуктивная	Тёр	Один компонент обозначает исходную точку для воспроизведения или превращения для другого компонента
Ситуативная	TE	Один компонент обозначает ситуацию, определяющую состояние или область действия второго компонента
Трансгрессивная	Trg	Один компонент обозначает результат превращения второго
Финитивная	Fin	Один компонент имеет значение цели, назначения другого

Вместо понятия отношения Г.С. Осипов ввел понятие семантических связей. И если Д.А. Поспелов рассматривал около 200 базовых отношений, то Г. С. Осипов ограничился 17 семантическими связями [4]. Они собраны таблице 2.

Можно заметить, что все семантические связи, описанные в таблице 2, представлены в медицинских текстах. Однако, наиболее часто встречаются каузальные (например, лучевая болезнь является причиной появления язвенных образований на коже), комитативные (например, грипп сопровождается высокой температурой), коррелятивные (например, грипп может привести к нарушению двигательных функций организма человека), потенсивные (например, гастрит увеличивает вероятность появления в будущем язвы желудка), результативные (например, цирроз печени может явиться следствием употребления определенных лекарств).

Выводы

Из анализа существующих методик, разработанных для «понимания» текста на русском языке, следует, что после модификаций их можно применить для «понимания» медицинского текста на русском языке, так необходимого для автоматизированного формирования базы знаний для медицинской экспертной системы.

Литература

1. Шенк Р. Обработка концептуальной информации. - М.: Энергия, 1980. - 360 с.
2. Логический подход к искусственному интеллекту: От модальной логики к логике баз данных: Пер. с франц./Тейз А., Грибомон П., Юлен Г. и др. - М.:Mnp, 199S.- 494с.
3. Поспелов Д. А. Логико-лингвистнческие модели в системах управления. - М.: Энергоиздат, 1981.-232 с.
4. Осипов Г. С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. - М.: Наука. Физматлит, 1997. - 112 с.