Естественный язык для схематизированных областей
(об информационной системе по анатомии W. Hagamen'a).

Владимир Борщев, ВИНИТИ РАН
Barbara H. Partee, University of Massachusetts, Amherst

1. Введение

У этой работы две части, или, может быть, лучше сказать, две темы. Первая – это описание одной конкретной диалоговой системы. Систему эту создал профессор анатомии Корнельского университета W. Hagamen (Hagamen 1998, 2002.), чтобы помочь студентам систематизировать их знания. По сути дела – это представление в компьютере Мира Анатомии и реализация диалога о сущностях этого Мира. С формально-терминологической точки зрения можно сказать, что это база данных по анатомии, в которой представлены анатомические факты – все кости, мышцы, нервы и другие «объекты» человеческого тела, их функции и соотношения – какая мышца что двигает, какие нервы и кровеносные сосуды ее «обслуживают» и т.п. Кроме того, над этой базой данных есть несколько очень интересных «надстроек».[1]

При этом общаться с системой можно на чистом английском языке, в частности, задавать ей вопросы, например:

(Q1) What muscle arising from the humerus that inserts on the radius, flexes the forearm and is innervated by the superficial radial nerve?

И она ответит:

(A2) The brachioradialis.

Более того, даже если в этом вопросе заменить некоторые корни абракадаброй, например, сочетанием xxx, сохранив окончания (см. Q3 ниже), т.е. превратить этот вопрос частично в «глокую куздру», система даст тот же ответ A2, который, как утверждает Hagamen, даст и любой человек, хоть как-то знающий анатомию:

(Q3) What muscle xxxing from the humerus and flexing the xxx is xxxed by the radial nerve?

Системе можно задавать и так называемые Why-questions, например Q4:

(Q4) Why does the pronator teres flex the forearm?

Вопрос этот толкуется так: нет ли таких закономерностей, из которой следует соответствующий факт? И система пытается такие закономерности найти. В данном случае она ответит:

(A5) All muscles that originate from the medial epicondyle of the humerus flex the forearm. All muscles that are anterior to the elbow joint flex the forearm.

Кроме того система и сама может задавать контрольные вопросы и проверять, правильно ли студенты на них отвечают. А в случае неправильных ответов, задавать более простые вопросы, выясняя, чего именно не знает студент.

Система представляется нам очень интересной, особенно ее диалоговая часть, реализующая некоторый фрагмент английского языка.

Этот фрагмент естественного языка – наша вторая тема, для нас еще более интересная: какие средства естественного языка используются в системах такого типа и вообще, в

[1] Наше описание системы несколько отличается от авторского, достаточно своеобразного. Мы пользуемся привычными нам понятиями информатики, логики и лингвистики, в надежде, что нам удается не искажать суть дела.

ситуациях, когда мы говорим о точных вещах. Тема эта далеко не нова. Из отечественных работ можно вспомнить, например, начатые еще в конце 50-х годов исследования А.В.Кузнецова, Е.В.Падучевой и Н.М.Ермолаевой (Кузнецов 1961) по описанию языка геометрии. Сравнительно недавно один из авторов писал в НТИ (и рассказывал на Диалоге-95) об интересных работах А.Б.Сосинского, предложившего очень простой фрагмент языка для математических текстов (Борщев 1994). .

Тут возникает несколько вопросов. Прежде всего, действительно ли мы имеем здесь дело с естественным фрагментом естественного языка, а не с некоторым фиксированным набором вопросов и ответов? И если это на самом деле фрагмент, то какие свойства выделяют его из всего языка, в чем его особенности, чем он лингвистически интересен?

Ответы на оба эти вопроса интересны и, во многом, далеко не тривиальны. Начнем с первого. Язык, на котором система разговаривает со своими пользователями, вопросы, которые она воспринимает и ответы, которые она выдает (и наоборот, вопросы, которые она задает, и ответы, которые она воспринимает) на наш взгляд, несомненно является фрагментом естественного языка. В этом утверждении есть две части. Во-первых, набор таких вопросов в каком-то смысле неограничен, и при этом они являются нормальными предложениями естественного языка. Во-вторых, система их в некотором смысле действительно «понимает», т.е. сопоставляет им правильную семантику, что в данном случае легко верифицировать – она совершает соответствующие операции над отношениями базы данных. В этом нет никакого чуда – базисные вопросы, понимаемые (и задаваемые) системой, эквивалентны, грубо говоря, формулам логики предикатов (точнее, некоторому подклассу таких формул), а семантика языка логики предикатов хорошо известна. Достижение (далеко не тривиальное) состоит в том, что в системе используются не формулы, а предложения естественного языка.

В ответе на второй вопрос тоже есть две части – какими чертами естественного языка этот фрагмент обладает, и каких черт и возможностей естественного языка в нем нет. Фрагмент этот достаточно богат, мы обсуждаем его в работе. Упомянем пока только одну его особенность – система широко использует типы (сорта) объектов.

В то же время, фрагмент этот, конечно, использует весьма ограниченную часть средств естественного языка. Это естественно и это ни в коей мере не является недостатком системы. Важно понять, где проходит эта граница – это одна из задач работы.

Язык велик и могуч. Обычно, описывая какую нибудь конкретную ситуацию, мы прежде всего структурируем, схематизируем ее средствами языка. Эти структуры, схемы мы как бы накладываем на описываемый фрагмент реальности (или выявляем их в нем). Причем мы делаем это в процессе построения текста, обладая свободой выбора – как именно структурировать и концептуализировать эту реальность. Как правило, мы можем делать это разными способами.

В данном же случае, как и всегда, когда мы говорим о вещах точных, структура, схема фрагмента реальности нам дана заранее, известна нам и нашему собеседнику. И мы, описывая реальность, следуем этой схеме. Это упорядочивает и, тем самым, упрощает семантику текста (дискурса). Набор объектов фиксирован и их имена, при некотором разнообразии, как бы известны заранее, грубо говоря, это «ярлыки», соответствующие элементам схемы. Набор отношений системы тоже фиксирован и только этими отношениями мы пользуемся. Упрощенная семантика, в свою очередь, дает возможность упрощать и упорядочивать другие средства языка.

Система Hagamen’a дает повод поговорить об этом четче и подробнее. К сожалению, жесткие ограничения на объем данного текста вынуждают нас к не слишком формальному и, при этом, сжатому изложению, а иногда делает его пунктирным. Увы, это относится в

наибольшей мере к самой интересной части системы – к ее «языковой» оболочке и к особенностям использованного фрагмента естественого языка.[2]

2. Мир Анатомии

Он состоит из *объектов* и *отношений*, связывающих эти объекты. Объекты разделены на *типы*: **кости**, **мышцы**, **вены**, **артерии**, **нервы** и т.п. Объекты имеют, как правило, канонические имена – принятые в анатомии названия. В системе, естественно, используются, имена, принятые в англиоязычной анатомии, и преставляющие исторически сложившуюся смесь английского с латынью: ACROMION, BRACHIORADIALIS, DEEP BRACHIAL ARTERY и т.д. В анатомических текстах и в повседневной практике используются и «неканонические» имена, причем возможна омонимия. Все такого рода имена могут употребляться на входе и выходе системы, в вопросах и ответах. Для «внутреннего», системного представления все объекты «для порядка» перенумерованы, и можно сказать, что там они представлены своими номерами. В терминологии баз данных такого рода «внутренние» представления объектов называются их *суррогатами*. Но наглядности ради мы будем игнорировать это внутреннее «числовое» представление объектов и считать, что объекты представлены в системе своими каноническими именами.

Отношения рассматриваются только бинарные, причем это могут быть как отношения между объектами разных типов, так и между объектами одного и того же типа. Отношения тоже разбиты на группы: **прикрепления** (**attachments**), **действия** (**actions**), **части** (**parts**) и т.п. Приведем только некоторые примеры.

«Прикреплений» всего два. Это отношения **Originate from** и **Insert on**. Каждая мышца где-то «начинается» (**Originate from**), обычно на какой-нибудь кости или ее части, и куда-то «прикрепляется» (**Insert on**) – тоже, как правило, к какой-нибудь кости или к участку кожи.[3]

Так, например, уже упоминавшаяся мышца BRACHIORADIALIS связана отношением **Originate from** с плечевой костью (HUMERUS) и отношением **Insert on** с лучевой костью (RADIUS).

«Действий» много. Мышцы что-нибудь (в основном, кости) сгибают (**Flex**), «приводят» (**Adduct**), «отводят» (**Abduct**) и т.п. Нервы что-нибудь (в основном, мышцы) иннервируют (**Innervate**), артерии «снабжают» (**Supply**) и т.д.

«Части» – это собственно отношение «быть частью» (**Is part of**) на костях и на мышцах и сходное с ним отношение «быть ветвью» (**Is a branch of**) на нервах, артериях и венах.

Есть и другие отношения, описывающие, например, разного рода пространственные соотношения объектов и пр.

Для отношений, так же, как и для объектов, фиксированы их канонические имена. Но в вопросах и ответах (и, видимо, вообще в анатомических текстах), используются и неканонические имена, причем и здесь эти имена часто неоднозначны, присутствует и синонимия, и омонимия. Отношения тоже перенумерованы и представлены в системе этими номерами.

Нам было дано полное описание верхней конечности. Там выделен 381 объект, объекты разделены на 15 типов. Рассматривается 25 различных отношений, содержащих 1130 «базисных фактов» (строк в таблицах отношений – см. примеры ниже).

---

[2] Полную версию данной работы предполагается опубликовать в сборнике «Научно-техническая информация».

[3] Чем отличается «начало» мышцы от ее «конца» не так легко определить, но анатомы пользуются некоторыми критериями, которые нет нужды здесь описывать; важно, что во всех случаях это четко определено и система следует этим представлениям анатомов.

3. Реляционные базы данных и логика предикатов

Представление фрагментов действительности как множеств объектов и набора отношений на этих объектах используется в реляционных базах данных. Поэтому нам удобно будет описывать представляемый в системе Мир Анатомии как реляционную базу данных.[4]

Заметим, что обычно в базах данных говорят о состояниях, каждое состояние представляет изображает состояние соответствующего фрагмента действительности в данный момент времени. Но Мир Анатомии не меняется во времени и в нашей базе данных только одно состояние, представляющее этот Мир.

В реляционных базах данных принято изображать отношения в виде таблиц. Примеры (для нашего случая) приведены ниже:

T1

Originate from	
что	**откуда**
FLEXOR POLLICIS LONGUS	ANTERIOR SURFACE OF THE RADIUS
LATERAL PART OF THE DELTOID	ACROMION
…	

T2

Insert on	
что	**куда**
TRAPEZIUS	ACROMION
PRONATOR QUADRATUS	ANTERIOR SURFACE OF THE RADIUS
…	

Объекты системы и их принадлежность к тем илии иным типам можно представить в виде унарных отношений:

T3

Muscle
ABDUCTOR DIGITI MINIMI
ABDUCTOR POLLICIS BREUIS
…

T4

Bone
ACROMION
ANTERIOR_SURFACE_OF_THE_ RADIUS
…

Эти отношения представляют базисные факты, содержащиеся в системе. На их основе строится «вопросно-ответная система», которой, по сути дела, является всякая база данных. Действительнно, базы данных создаются не только для того, чтобы хранить данные, но и

[4] Подробнее о реляционных базах данных можно прочесть, например, в книге Дейт (1980) или в статье Борщев (1982).

чтобы получать их оттуда в самой разнообразной форме, в ответ на соответствующие *запросы*.

Реляционные базы данных основаны на логике предикатов. Язык логики используется обычно и как основа языка запросов, и для описания законов моделируемого фрагмента реальности, прежде всего – свойств отношений.

Мы не будем здесь описывать логику предикатов, ограничимся «анатомическими» примерами и неформальными пояснениями. Конкретный язык – в данном случае «анатомический» – будет фиксирован, если мы фиксируем его основные кирпичики, простейшие *выражения – предметные константы* (имена объектов), *предметные переменные* и *предикатные символы* (имена отношений). В данном случае в качестве предметных констант мы будем использовать канонические имена анатомических объектов, а в качестве предикатных символов – канонические имена отношений и типов объектов (для унарных отношений, педставляющих эти типы). Из перечисленных выше кирпичиков строятся *атомарные формулы*, например:

(F1) **Originate from** (BRACHIORADIALIS, HUMERUS)

(F2) **Insert on** (***y***, RADIUS)

(F3) **Muscle** (***y***)

(F4) **Nerve** (SUPERFICIAL RADIAL NERVE)

Из атомарных формул с помощью логических связок и кванторов строятся все другие формулы. Например:

(F5) **Muscle**(***y***) & **Originate from** (***y***, HUMERUS) & **Insert on** (***y***, RADIUS) & **Flex** (***y***, FOREARM) & **Innervate** (SUPERFICIAL RADIAL NERVE, ***y***)

(F6) $\forall x \forall y$ (**Innervate** (***y***, ***x***) $\rightarrow$ **Nerve** (***y***))

В формулах могут содержаться *свободные переменные* – в F2 это ***x***, а в F3 и F5 – это ***y.*** В F1 и F4 вообще нет переменных, а в F6 все переменные *связаны* кванторами. Такие формулы называются *замкнутыми*.

Внимательный читатель заметит что формула F5 соответствует вопросу Q1 выше, а формула F6 – это запись очень простого анатомического «закона» – *только нервы иннервируют*. Системе, чтобы быть эффективной, нужно учитывать массу таких простейших закономерностей.

Формулы *интерпретируются* на *моделях*. В реляционных базах данных моделями, служат состояния базы данных. В нашей анатомической базе данных ровно одно состояние – это описанный выше Мир Анатомии, анатомические объекты и отношения на них, т.е. приведенные выше таблицы. При реализации базы данных эти объекты и таблицы должны быть представлены как-то в памяти компьютера.

Каждая замкнутая формула истинна или ложна в данной модели. Истинность атомарных формул задается интерпретацией предикатных символов, скажем, формула F1 верна в нашей модели, т.к. строка BRACHIORADIALIS, HUMERUS содержится в таблице **Originate from**. Истинность сложных формул вычисляется по правилам интерпретации логических связок и кванторов и значениям атомарных формул.

Каждой незамкнутая формула задает на модели отношение, *производное* от базисных отношений модели. Размерность (арность) этого отношения определяется числом свободных переменной формулы.

Как мы уже говорили, формулы могут рассматриваться как запросы к базе данных. Если формула замкнута, то ответом будет *да*, если формула истинна, и *нет* в противном случае.

Ответом на запрос, соответствующий незамкнутой формуле, будет производное отношение (таблица), сооветствующая данной формуле.

По сути дела, эту новую таблицу можно получить из старых, базисных, если их «резать» и «клеить» в соответствии с формулой-запросом. Так, чтобы ответить на вопрос (1), или, что то же самое, запрос-формулу F5, система выберет из таблицы T1 все мышцы, «начинающиеся» на кости HUMERUS, а из таблицы T2 все мыщцы, «прикрепленные» к кости RADIUS. Кроме того, надо посмотреть в таблицы отношений **Flex** и **Innervate** и найти там мышцы, соответственно, сгибающие предплечье и иннервируемые поверхностным лучевым нервом (SUPERFICIAL RADIAL NERVE), а потом взять те из них, которые содержатся во всех четырех выборках из этих четырех таблицах (т.е теоретико-множественное пересечении этих выборок). Оказывается, что есть только одна такая мышца – BRACHIORADIALIS, т.е. в данном случае таблица-ответ состоит из одной строки. Но вообще говоря, таблица-ответ может состоять и из многих строк, а может быть пустой.

3. Языковая оболочка

3.1. Для чего нужен был предыдущий раздел.

Как мы уже говорили, пользователи общаются с системой Hagamen'a на английском языке. Зачем же мы рассматривали логику, как язык запосов к базе данных? Дело в том, что задавая вопрос системе, мы просим ее произвести четкие действия над формальными объектами – некоторые операции над отношениями, хранящимися в системе. И с точки зрения этих операций вопросы, на которые умеет отвечать система, эквивалентны некоторому подклассу формул логики предикатов (т.е. ответы на них получаются по правилам, соответствующим семантике этих формул).[5]

3.2. Как система анализирует вопросы на естественном языке.

Из того, что система рассматривает вопросы, эквивалентные формулам логики первого порядка, вовсе не следует, что анализ этих вопросов – вещь тривиальная. Отнюдь. Синтаксис таких вопросов может очень сильно отличаться от синтаксиса логических формул.

Мы опишем ниже работу системы в самых общих чертах, крупными блоками и на примере. Рассмотрим вопрос Q1' (упрощенную версию приведенного выше вопроса Q1):

(Q1') What muscles arising from the humerus flex the forearm?

Вопрос этот, как нетрудно видеть, соответствует формуле F5':

(F5') **Muscle** (***y***) & **Originate from** (***y***, HUMERUS) & **Flex** (***y***, FOREARM)

Грубо говоря, анализирующая программа, используемая в системе, выполняет следующие работы (речь идет не о последовательности этих задач, а о их сути):

(1) «Разрезает» вопрос на «блоки», соответствующие именам объектов и отношений.

(2) Отождествляет эти блоки (с помощью соответствующих словарей) с объектами и именами отношениями.

(3) Строит синтаксическую структуру, связывающую имена объектов с именами отношений (т.е., фактически, аналогичную структуре соответствующей формуле).

(4) Интерпретирует эту структуру в базе данных (т.е. выполняет нужные операции над таблицами отношений) и, тем самым, получает ответ на вопрос.

Не вдаваясь в детали, поясним эти действия на нашем примере Q1'. Чтобы выполнить пункт (1) система проводит грубый синтаксический анализ, на основе так называемой

[5] Это не относится к вопросам типа «Почему», с которыми все чуть сложнее.

«грамматики границ» (обращаясь к словарям,учитывая предлоги, окончания, etc). И получает нечто вроде Q2”:

(Q”) What muscles | arising from | the humerus | flex the forearm

Для выполнения пункта (2) нужны словари, представления о структуре имен и их типов, для пункта (3) – допущения о синтаксической структуре, т.е. грамматика в той или иной форме (заметим, что синтаксическому анализу в данном случае подвергается не цепочка слов, а результат предыдущих этапов, т.е. цепочка «блоков» – имен объектов и отношений). Результатом этапа (3) является, по сути дела, формула и этап (4) аналогичен интерпретации этой формулы.

4. Рассматриваемый фрагмент и его место в языке

Читатель, мы надеемся, убедился в том, что в этой системе мы действительно имеем дело с фрагментом естественного языка. Но чем этот фрагмент интересен? Что в нем есть и чего в нем нет? И какие его особенности используются в такого рода системах?

4.1. Четко структурированный фрагмент мира

О главном отличии уже говорилось во введении. Прежде всего, мы имеем здесь дело с фрагментом действительности, структура которого уже определена, фиксирована в системе, и, как правило, известна пользователю этой системы (если он даже и не знает этой структуры полностью, он с ней знакомится на примерах, но не в силах ее изменить).

Более того, в данном случае структура это достаточна простая и может быть представлена как набор множеств и отношений на них, т.е в виде реляционной базы данных.

Обычно же, пользуясь языком даже в самых простых ситуациях, скажем описывая какую-нибудь «картинку» – что, например, сейчас находится у нас на столе или на книжной полке – мы структурируем ситуацию сами и не связаны никакими ограничениям, можем самыми разными способами характеризовать находящиеся там объекты и расматривать любые подходящие случаю связи и отношения между ними.

4.2. Система ограничений

Конкретный фрагмент языка, используемый при общении с данной системой, удовлетворяет целому ряду конкретных ограничений.[6] Рассмотрим только одно из них.

Никаких модификаторов. Это очень важное ограничение. Кау уже указывалось, все объекты и отношения системы заданы соответствующими списками. У них бывают канонические и неканонические имена, возможна омонимия и синонимия. В процессе анализа этих имен омонимия разрешается, объекты и отношения, так сказать идентифицируются, но, грубо говоря, никакой семантической структуры этих имен не предполагается, имена понимаются как ярлыки (как «компаунды»).

Внешне имена анатомических объектов не отличаются от сложных имен в естественном языке. Расмотрим, например, каноническое имя одной из артерий:

ASCENDING BRANCH OF THE DEEP BRACHIAL ARTERY

Когда мы встречаем такую именную группу в тексте, мы понимаем, что речь идет о ветви конкретной артерии – DEEP BRACHIAL ARTERY, причем это ветвь восходящая (ASCENDING), видимо, есть и другие, не восходящие. А сама артерия находится где-то в глубине (DEEP), видимо есть и другие, более поверхностные. Т.е. мы устанавливаем сразу несколько отношений между реалиями.

---

[6] Надо отметить, что сам Hagamen никаких ограничений не формулирует. Но их можно извлечь из примеров и описания работы системы.

А вот система Hagamen'a, как и все системы такого рода, не наделенные тонким семантическим анализом (который при существующем состоянии семантики нелегко себе представить), воспринимает данное сложное имя просто как сложный ярлык, сложное собственное имя, и соотносит его с соответствующим объектом в словаре, *не устанавливая никаких отношений* с другими объектами. Даже тот факт, что данный объект есть ветвь (часть) другого объекта (артерии), при том, что отношение быть ветвью (**is a branch of)** входит в список отношений системы, не будет установлен при анализе имени, а может быть установлен только по таблице отношений.

Таким образом, фрагмент естественного языка (рассматриваемый с точки зрения системы) есть нечто сравнительно примитивное по сравнению со всем языком даже в его, так сказать, базисных функциях.

Подчеркнем, что оценочность данного высказывания не есть оценка системы, а просто констатация сложности естественного языка. Построение такого рода систем позволяет почувствовать эту сложность и искать пути усовершенствования этих систем.

Литература

Борщев В.Б. (1994) Формальный язык как часть естественного, *НТИ*, Серия 2, № 9, 1994, 27-31.

Борщев В.Б. (1982) Базы и банки данных, *Природа*, № 3, 1982, 64-75.

Дейт К. (1980) *Введение в системы баз данных*. М: Наука, 1980.

Кузнецов А.В., Е.В. Падучева и Н.М. Ермолаева (1961). Об информационном языке для геометрии и алгоритме перевода с естественного языка на информационный и обратно. В. сб. *Лингвистические исследования по машинному переводу*. М.. 1961.

Hagamen, W.D. (1998) Anatomy of meaning. Ms.

Hagamen, W.D. (2002) Meaning Units and the Natural Language Semantics of Anatomy. Ms.