

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ЗНАНИЙ О ТАКСОНОМИЯХ ИЗ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

И.С. Мошков

Самарский государственный технический университет
443100, г. Самара, ул. Молодогвардейская, 244

Рассмотрены синтаксическая и семантическая структуры текстов таксономического характера на естественном языке. Для этого проведен системный анализ лексики текстов и определена связь основных лексических конструкций с их значением. На основе данного анализа строится модель субъективных знаний о таксономии, описанной в тексте, позволяющая автоматизировать процесс извлечения из текста описываемой в нем таксономической структуры.

Ключевые слова: автоматизация, знания, естественный язык, таксономии.

Извлечение информации из текстов на естественном языке (ЕЯ) востребовано в различных областях [1, 2, 3]. Кроме того, практически в любой области имеется тенденция к систематизации накопленных знаний [4]. При этом анализ текста и построение субъективной модели классификации, описанной в тексте, на основе проведенного анализа позволяют дать сравнительную оценку эталонных и субъективных знаний о предметной области [1]. В дальнейшем сравнительная оценка может использоваться как один из этапов «понимания» текста на ЕЯ. Однако приходится извлекать информацию из неформализованного текста, который обладает свойством избыточности, а также может содержать допущенные ошибки и осколочные фразы, что приводит к значительному усложнению алгоритмов анализа [1, 5]. Следовательно, необходимо иметь максимальное количество информации об основных принципах построения высказываний на ЕЯ.

Так как знания в своей основе задаются оппозициями, то есть определяется некоторое значение предметной области путем противопоставления другим значениям, то представление знаний в виде таксономий является одной из наиболее наглядных форм. С другой стороны, оценка таксономических знаний человека является важным фактором оценки его компетентности, так как можно проследить зависимость между накопленными знаниями о предметной области и умением качественно их структурировать [1, 4, 6]. Задача автоматического анализа качества субъективных таксономий является актуальной из-за того, что позволяет сократить время контроля имеющихся знаний о предмете, так как в настоящее время наиболее распространенным методом оценки компетентности являются тесты закрытого типа и метод экспертной оценки. При этом можно ограничиваться текстом, описывающим таксономию. Это означает, что, с одной стороны, можно не вводить искусственного языка для высказываний человека, с другой стороны, можно оценить знания, содержащиеся в тексте, в достаточной мере.

Для того чтобы распознавать классификации, необходимо выделить ряд особенностей, характерных для них, которые выражаются в тексте на ЕЯ. Во-первых, любая классификация состоит из элементов и сложность описания классификации зависит от сложности описания элемента. Во-вторых, элементы связаны между собой, поэтому в тексте также будет встречаться и описание связей элементов. Кроме этого,

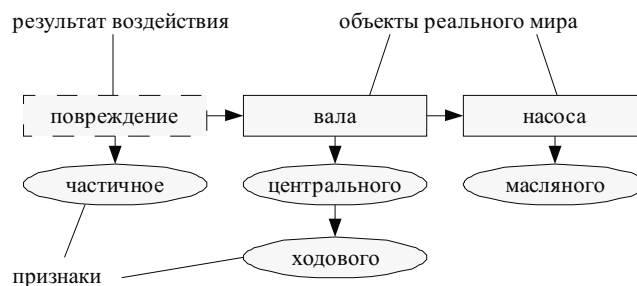
в зависимости от сложности объекта, задаваемого классификацией, зависит его описание – чем больше в нем элементов, тем больше в тексте будет встречаться описание неименованных элементов, составленных из ряда более простых [1].

В языке принято выделять «слои» языка: слой языка, описывающий объективную реальность (язык-объект) и слой языка, описывающий действия с языком-объектом (объектный метаязык). В структуре языка также принято выделять синтагматический и парадигматический уровни организации [7]. Считается, что синтагматический уровень определяет способ связи слов в высказываниях, а парадигматический отражает систему значений (семантику) предметной области. Следовательно, задача анализа лексики, типичной для таксономических текстов, сводится к разработке метода, распознающего особенности описания различных слоев языка в тексте. Далее слой языка объекта используется для создания вершин таксономической структуры, а объектный метаязык задает ее ребра.

Анализ текстов, описывающих таксономическую структуру, показал, что их лексика, относящаяся к слою языка объекта, делится на три больших и относительно самостоятельных группы:

- 1) группа имен существительных, относящихся к обозначению объекта реального мира или его составных частей (группа «сущности»);
- 2) лексика, описывающая признаки основных значений данной предметной области. В речи выражается в основном именами прилагательными, реже – предложно-падежными конструкциями, согласованными с именами существительными (группа «признаки»);
- 3) группа имен существительных, описывающих воздействие на сущность или результат этого воздействия (группа «операторы»).

Пример деления лексики, описывающей смысловой объект, приведен на рис. 1.



Р и с. 1. Пример структуры смыслового объекта

К слою объектного метаязыка была отнесена лексика, определяющая логические операции с объектами, описываемыми уровнем языка-объекта:

- 1) предлоги, задающие связи между различными объектами слоя языка-объекта;
- 2) лексика, описывающая критерии и способы деления.

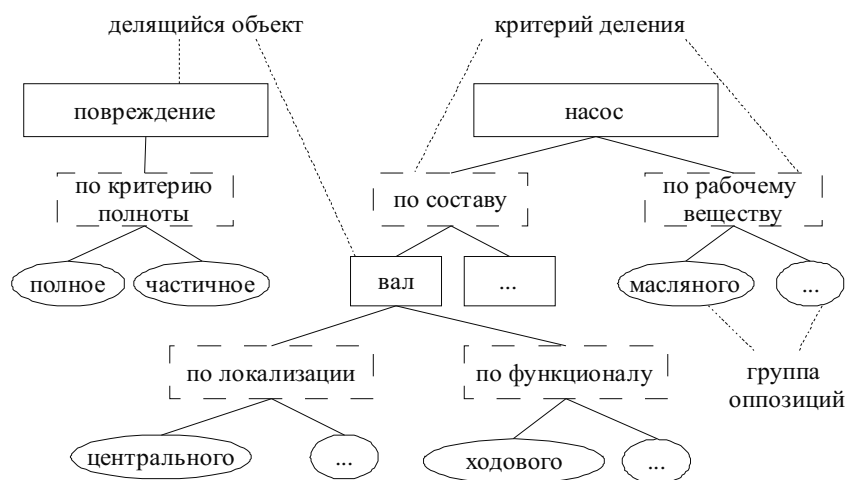
Выделенные лексические группы позволяют извлечь информацию с максимальной детализацией в случае разложения текста таксономического характера. В случае, когда в тексте присутствуют сложные обороты, несущие дополнительную информацию, возможно сведение неклассифицированной лексики к данным группам с потерей детализации, но без потери информации. Такой подход позволяет анализировать сложную информацию путем заложения часто встречаемых сложных оборотов в базу знаний. Впоследствии можно, используя дополнительные алгоритмы разбора, сделать распознаваемую структуру более детализированной. Также следует отме-

титель, что если лексика языка-объекта зависит от предметной области, то лексика описания связей от нее практически не зависит и имеет сравнительно меньший объем. Это позволяет построить набор команд, оперирующих с деревом, которые будут изменять структуру дерева в зависимости от смысла слова из лексики связей.

На основе особенностей лексики можно определить специфику базы знаний, в которой будут храниться смысловые значения используемой лексики (рис. 2):

- таксономическая структура, задающая общепринятую классификацию деления смысловых объектов (эталон), должна содержать два вида вершин: критерии деления и элементы деления;

- для каждого класса таксономии смысловых объектов необходимо задавать собственное дерево классификации. По типу такие деревья делятся на признаковые и объектные. В большинстве случаев возможность синтагматической связки слов в тексте определяется классом таксономии, к которой относится тот или иной элемент.



Р и с. 2. Пример имеющейся структуры классификации объектов в базе знаний

Зададим множество значений предметной области $T = \{t_1, \dots, t_n\}$, где t – значение предметной области, заданное набором его возможных представлений в тексте. Определим элемент e , который является некоторым значением предметной области, и предикат $Sub(e_i, e_j)$, который определяет, является ли e_i подвидом e_j или его составной частью. А также определим множество E_k , которое включает все непосредственные подвиды для элемента e_k , то есть

$$E_k = \{e_1, \dots, e_n\}, \forall e_i \in E_k \Rightarrow Sub(e_i, e_k) = 1. \quad (1)$$

Таким образом, множество $K = \{E_1, \dots, E_n\}$, включающее все элементы e и отношения между элементами, будет определять простейший вид классификации. Пример: «Стул состоит из ножек, спинки, сиденья. Ножка стула состоит из <...>». В данном случае для определения каждого из элементов используются совершенно независимые термины. Если в качестве связи между элементами используется отношения «часть – целое», то каждый следующий уровень в тексте обозначается как набор терминов, указывающих путь в используемой классификации. Таким образом,

в данном случае субъективная семантическая модель $Sem = \langle K_0 \rangle$ состоит из одного дерева классификации $K_0 = \{E_1, \dots, E_n\}$ элементов $e = \{t_1, \dots, t_n\}$.

Далее рассмотрим вариант, когда элемент e состоит из подэлементов s и P , где $s = \{t_1, \dots, t_n\}$ – описываемый субъект, а P – множество признаков $P = \{p_1, \dots, p_n\}$, задающих отдельное свойство субъекта. Тогда для определения нового элемента можно менять как признаковую часть, так и термин субъекта. Пример: «По химической классификации нефть делится на парафиновые нефти, нафтенковые нефти, ароматические нефти». Данный вариант описания классификации является наиболее часто используемым. Таким образом, субъективная семантическая модель будет описываться как $Sem = \langle K_s K_p K_0 \rangle$, состоящая из трех деревьев, каждое из которых описывается как $K_i = \{E_1, \dots, E_n\}$. Элементы дерева описываются как $e = \langle s, P \rangle$, где $P = \{p_1, \dots, p_n\}$ и $s = \{t_1, \dots, t_n\}$.

Также есть вариант сложной классификации, когда в одном элементе присутствуют простые элементы из различных классификаций. В данном типе классификации элемент e в общем случае состоит из подэлементов s , P и подэлемента-объекта o , который характеризует область приложения s . Объект o в свою очередь может иметь такую же структуру, как и весь элемент e . Пример: «Повреждения рельсов делятся на изгибы, повреждения в шейке, изломы по всему сечению и дефекты подошвы. Изломы бывают поперечными с видимыми пороками и без видимых пороков». В приведенном примере выделен класс «повреждения», куда входят элементы «изгибы», «изломы», «дефекты», а также класс «рельс», куда входят элементы «шейка», «полное сечение», «подошва». Также элементы классификации «повреждений» оперируют с элементами класса «рельс». Данный вариант построения классификации является наиболее общим и формально может включать неограниченное число используемых классификаций при построении элемента, однако на практике редко используется более двух классов. Таким образом, для распознавания такого вида классификаций субъективная семантическая модель будет описываться как $Sem = \langle \{K_1, \dots, K_n\}, K_0 \rangle$, где присутствует множество деревьев классификаций простых значений ($o_{simple} = \langle s, P \rangle$) и общее дерево классификации значений предметной области, состоящей из комплексных значений предметной области $o_{complex} = \langle s, P, o \rangle$.

Помимо всего прочего таксономические конструкции могут иметь более сложную структуру. В них могут встречаться структурные элементы, в которых один элемент делится на другие элементы по разным критериям деления. Поэтому введем предикат $Crit(e, K_s)$, который определяет, входит ли элемент e в критериальное подмножество K_s . Причем один элемент может входить в несколько критериальных подмножеств. Таким образом, более полным описанием структуры одной классификации является $K = \{K_{s_1}, \dots, K_{s_n}\}$, где за исключением e_0 – корневой вершины дерева классификации – будет выполняться условие

$$K_{s_i} = \{e_1, \dots, e_n\}, \forall e_j \in K_{s_i} \exists e_i | Sub(e_i, e_j) \cap Crit(e_j, K_{s_i}) = 1. \quad (2)$$

Следует отметить, что для русского языка характерна неполнота, поэтому при анализе классификаций возникает задача восстановления встречающихся элементов

до полного вида [5]. При этом следует опираться на выделенные значения предметной области родительских элементов.

Дальнейшее развитие сложности классификаций приводит к тому, что растет количество составных частей элементов, однако структура элемента классификации остается практически неизменной. Например, признаковая часть может выражаться как прилагательными, так и причастными оборотами. Поэтому нахождение объектов в тексте на основе данной структуры является достаточным для широкого ряда задач. Выделять элементы из высказывания на ЕЯ можно на этапе синтаксического анализа на основе ролей слов в предложении и затем пополнять на этапе семантического анализа, когда определяется отношение между словами и фразами. Таким образом, на основе приведенной модели можно проводить анализ текста на ЕЯ, содержащий описание классификации любой предметной области, путем выделения из текста элемента классификации, их отношений и сравнения полученных моделей с эталонными.

Построение таксономических текстов характеризуется описанием семантики предметной области и языковым выражением этого описания. В настоящее время основное внимание уделяется построению системы терминов и анализу лексики. Однако лингвисты показывают, что необходимо анализировать высказывания, состоящие из отдельных предложений, как целостную систему смыслов [5]. При этом в доступной литературе отсутствует описание единого методологического подхода к решению данной проблемы. Поэтому была поставлена задача анализа структуры классификаций как в технических, так и в естественных науках с точки зрения языка с тем, чтобы выявить основные закономерности построения этих таксономий, построить на этой основе базы знаний, провести анализ таксономических текстов и разработать формальную модель для автоматической оценки качества построения классификаций.

Традиционно в системе языка по качеству системной организации выделяют следующие языковые уровни: уровень морфем, уровень лексем, уровень словосочетаний (или составных терминов), уровень предложений, уровень абзаца и уровень текста в целом [2, 3]. Для того чтобы проанализировать весь текст, необходимо пройти ряд этапов анализа каждого из уровней.

В зависимости от анализируемого уровня и качества его анализа можно выделить следующие основные этапы: морфологический анализ, поверхностный синтаксический анализ, глубинный синтаксический анализ, семантический анализ, прагматический анализ [2, 5]. Поэтому задача анализа текста сводится к последовательному проведению этапов анализа текстов на естественном языке и получению таксономической структуры, описанной в тексте, в формальном виде.

Зададим входное высказывание как набор слов $Sen = \{W_1, \dots, W_n\}$, где n – число слов в высказывании. Когда на вход поступает набор слов, то следует сначала получить морфологическую информацию о каждом слове, поэтому зададим функцию F_m , которая сопоставляет некоторому слову набор морфологических характеристик $M_w = \langle \{W_0, \dots, W_6\}, \{W_{r_0}, \dots, W_{r_2}\}, \{W_{c_0}, \dots, W_{c_1}\} \rangle$, где $\{W_0, \dots, W_6\}$ – множество падежей, W_r – род, W_c – число. Данная морфологическая информация необходима на этапе синтаксического анализа – она позволяет определить зависимость между словами. Определим также соответствие между всеми возможными формами слова и формой заданного слова путем определения вектора $I_{w_i} = \langle P, R, C \rangle$, который для слова W_i указывает индекс характеристики формы. Таким образом, получаем новый тип данных, который описывает морфологические характеристики слова,

$W_m = \langle W, M_w, I \rangle$. Используя морфологические характеристики слова, можно переходить к этапу синтаксического анализа.

Определим общую синтаксическую структуру фразы, которая описывает смысловой объект. Для этого зададим простейший элемент объекта – S , который в общем случае является существительным; основным признаком этого элемента является отсутствие подчиненного слова. Также зададим корневой элемент O смыслового объекта, признаком которого является наличие подчиненного слова, которое может являться простейшим элементом S' или описываться такой же структурой O' , как и весь смысловой объект. Поэтому определим смысловой объект (СО) верхнего уровня как корневой СО. При этом как корневой элемент O , так и простейший элемент S могут иметь неограниченное количество признаков P , которые в основном могут описываться прилагательными, причастными оборотами, а также наречиями. По смыслу корневой элемент в смысловом объекте определяет основной объект, о котором идет речь, а элемент подчинения определяет субъект, который характеризует область приложения корневого объекта. Таким образом, получаем общую структуру смыслового объекта:

$$Obj = \langle P, O, Obj' \rangle, P = \{P_1, \dots, P_n\}; \quad (3)$$

$$Obj' = \begin{cases} S, Obj'' = \emptyset; \\ \langle P, O, Obj'' \rangle, Obj'' \neq \emptyset. \end{cases} \quad (4)$$

На синтаксическом уровне анализа производится попытка определить в высказывании границы всех смысловых объектов и построить связи между словами, входящими в один корневой объект. При этом приоритетным вариантом анализа является случай, когда одно предложение описывает некий смысловой объект и ряд подчиненных смысловых объектов с указанием связи между СО родителя и подчиненными СО. На выходе получаем набор смысловых объектов и слов, связывающих эти смысловые объекты. Далее проводится постсинтаксический этап анализа, на котором определяется смысл слов, задающих связи между смысловыми объектами, после чего получаем таксономическую структуру связанных смысловых объектов (особенности постсинтаксического анализа).

Таким образом, перед этапом синтаксического анализа на входе имеем множество предложений $\{\varphi_1, \dots, \varphi_n\}$, каждое из которых содержит множество слов $\{W_{i1}, \dots, W_{im}\}$. Так как еще нельзя связать отдельные фразы между собой, то на этапе синтаксического анализа определяются связи слов внутри фразы. Т. е. для фразы φ_k получаем $\langle \{W_{k1}, \dots, W_{kl}\}, \{\langle Obj_{k1}, Pos_1 \rangle, \dots, \langle Obj_{kp}, Pos_p \rangle\}$, где Pos_i – позиция корневого слова i -того СО в предложении k . Далее полученные сегменты собираются в одно синтаксическое дерево для всего таксономического текста и определяются типы связей между различными сегментами.

Значение (семантика) отдельного слова задается перечислением его возможных номинаций в тексте. В зависимости от типа передаваемого значения можно выделить несколько классов семантик, причем семантики, находящиеся внутри одного класса, заданы иерархической структурой отношений. Таким образом, имеется множество таксономий, которые задают отношения различных смыслов для отдельного класса. Значение сочетания слов, используемых для описания простого смыслового объекта, можно задавать несколькими способами. Например, путем создания эталонной таксономии, в которой задаются отношения между смысловыми объектами,

или путем задания семантических фильтров. Семантические фильтры разрешают или запрещают сочетание семантик на уровне слов или классов. Поэтому, имея синтаксическую структуру смысловых объектов, можно сопоставлять эталонные значения слов и оценивать корректность связей в эталонной модели и полученной субъективной таксономической структуре.

Таким образом, из текстов таксономического типа можно извлекать смысловые объекты и сложные группы смысловых объектов, информацию о критериях деления, метаязыковую информацию, описывающую структуру деления в тексте, и строить на основе этого граф, представляющий собой структуру, описанную таксономическим текстом, в различных предметных областях.

В заключение можно отметить, что задачи связи синтаксиса с семантикой и оценки семантики высказывания применительно к обработке результатов открытого тестирования являются решаемыми. Это позволяет отказаться от использования методов экспертных оценок и проводить развернутый автоматизированный процесс оценки качества высказываний, использующий квалифицирующую систему, со строгими и многофакторными критериями анализа ответов.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2000. – 384 с.
2. Селезнев К. Обработка текстов на естественном языке // Открытые системы. – 2000. – №12.
3. Попов Э.В. Общение с ЭВМ на естественном языке. – М.: Наука, 1982. – 360 с.
4. Никаев С.А. Модели и информационная система для оценки профессиональных знаний специалистов промышленного производства: Автореф. дисс. канд. спец. 05.13.01 Системный анализ, управление и обработка информации (промышленность). – Самара, 2004. – 24 с.
5. Шемакин Ю.И. Начала компьютерной лингвистики. – М.: Издательство МГОУ, А/О «Росвузнаука», 1992. – 81 с.
6. Солсо Р. Когнитивная психология. – 6-е изд. – СПб.: Питер, 2006. – 589 с.
7. Лурия А.Р. Язык и сознание / Под ред. Е.Д. Хомской. – Ростов н/Д.: Феникс, 1998. – 416 с.

Статья поступила в редакцию 29 сентября 2010 г.

UDC 519.767.6

AUTOMATIC ACQUISITION OF KNOWLEDGE ABOUT TAXONOMIES FROM FREE LANGUAGE TEXT

I.S. Moshkov

Samara State Technical University
244, Molodogvardejskaya str., Samara, 443100

The article examines syntactic and semantic structure of taxonomy-based natural language texts. In order to do this, texts vocabulary was analysed and connection between basic lexical structures and their meaning was determined. The analysis allows to construct a model of subjective knowledge about taxonomy described in the text. This model can be used for automated acquisition of taxonomic structure from the text describing it.

Keywords: automation, knowledge, natural language, taxonomy.

Iliya S. Moshkov – Postgraduate student.