disciplines (modules) according to a certain feature for the consequent transformation into work programs of disciplines are given.

Keywords: formalization of competences, automation of education plan development

Получено: 23.11.11

УДК 519.767.6

В. Н. Якимов, доктор технических наук, профессор
Самарский государственный технический университет
И. С. Мошков
Самарский государственный медицинский университет

СТРУКТУРНЫЙ АНАЛИЗ СЛОЖНЫХ ТЕРМИНОВ В ТЕХНИЧЕСКИХ ДОКУМЕНТАХ

Анализируются особенности текстов на естественном языке, описывающих таксономическую структуру. Основной упор сделан на классификацию элементов, из которых состоят термины в тексте. Также определены критерии принадлежности, по которым можно классифицировать тот или иной элемент сложного составного термина.

Ключевые слова: естественный язык, анализ текста, таксономическая структура

Введение

В соответствии с увеличением потока информации усложняются задачи автоматизации обработки данных, поступающих из различных текстовых документов. Поэтому актуальность разработки новых и совершенствования известных инструментов для извлечения информации из текста постоянно растет. Одним из способов применения данных инструментов является оценка знаний, содержащихся в тексте [1, 2], которая заключается в сравнении структуры знаний некоторого субъекта с эталоном и может использоваться как средство автоматической обработки результатов открытого тестирования [3]. Однако некоторые особенности текста на естественном языке (неполнота, избыточность, противоречивость) создают трудности в процессе создания инструмента для полноценного анализа текста [4]. Таким образом, возникает потребность в структурном анализе текстового представления информации и разработке формальных способов анализа текста, которые бы позволили, с одной стороны, проводить автоматический анализ текста, необходимого для оценки знаний, а с другой – упростить анализ за счет введения допустимых ограничений, сохраняющих необходимый уровень качества анализа. Одним из таких ограничений является использование в качестве анализируемого материала текста, описывающего таксономическую структуру. Это обусловлено тем, что практически в любой области науки и техники с точки зрения обеспечения системности требуется проводить структурирование и классификацию имеющихся знаний [5]. С другой стороны, существующие исследования показывают, что есть взаимосвязь между умением строить правильную классификацию понятий определенной предметной области и умением аргументированно принимать адекватные решения в данной предметной области [6, 7].

Анализ структурных особенностей текста, описывающего таксономию

Для того чтобы сформулировать требования к формальному аппарату анализа, поделим высказывание на естественном языке (ЕЯ), описывающее таксономию, на отдельные части и определим функции, которые они выполняют в тексте, а также возможные способы их нахождения. Ниже будем использовать высказывание φ, где φ – множество сложных составных терминов; L – связей между ними; K – критериев деления терминов; T – метаязыковых конструкций, описывающих качественные особенности таксономии. Для определенности в качестве примера будет использоваться высказывание: «По химической классификации нефть делится на три основные группы: парафиновые нефти, нафтеновые нефти, ароматические нефти».

В высказываниях на ЕЯ, описывающих таксономию, можно выделить четыре функциональных элемента:

1. *Obj* – описание элементов классификаций (сложных составных терминов – ССТ), которые в предложении, как правило, являются подлежащими и дополнениями, описывающими основную сущность термина, и согласованными с ними дополнениями и определениями, которые задают уточнения месторасположения элемента среди других элементов в таксономии. Таким образом, при условии нахождении главного слова ССТ уже на этапе синтаксического анализа можно определить его границы и структуру. ССТ являются основными элементами таксономии, причем существует некоторая «признаковая» часть, которая позволяет отделить значение одного ССТ от другого. Нахождение ССТ в тексте – сравнительно несложная задача, однако в русском языке, как правило, употребляются осколочные выражения: когда одна или несколько частей ССТ не описаны явно, а лишь подразумеваются. Поэтому поиск ССТ в высказывании усложняется и возникает дополнительная задача восстановления в высказывании полного описания ССТ. Для подтверждения правильности найденной части термина на этапе синтаксического анализа используется формальное описание предметной области экспертом (эталонная таксономия). В используемом примере это слова и словосочетания, называющие термины, относящиеся к классу нефтяных веществ: «нефть», «парафиновые нефти» и т. п.

2. *L* – описание связей между ССТ. В предложении обычно выражаются глагольной группой, а также системой падежей и знаков препинания. Поскольку в таксономии существуют два вида связей: «родитель – дочерний элемент» и «дочерний элемент – родитель», для распознавания связей можно использовать лексические шаблоны. При построении субъективной таксономии (таксономии, извлекаемой из анализируемого высказывания) данный элемент носит функциональный характер, связывая извлеченные термины. Существует также тип связей между ССТ, которые создают более сложный термин, выполняя операции сочетания двух ССТ. Причем один из терминов становится частью характеристики другого термина, соответственно эти термины могут не иметь видовой или композиционной связи. В используемом примере описанием связи является глагольная группа «делится на…».

3. *K* – критерии деления ССТ. В предложении обычно выражаются дополнениями. Критерии деления объясняют принцип деления и сопоставления различных элементов таксономии. В используемом примере критерием является словосочетание «по химической классификации».

4. *T* – описание каких-либо структурных особенностей таксономии. В используемом примере описанием структурных особенностей является словосочетание

«три основные группы», что означает намерение описания в субъективной структуре трех групп.

Термины, встречающиеся в высказывании, которое описывает таксономию, могут как находиться в отношении «вид» – «подвид», так и быть независимыми в данном отношении. На рис. 1 приведена структурная схема фразы используемого примера.

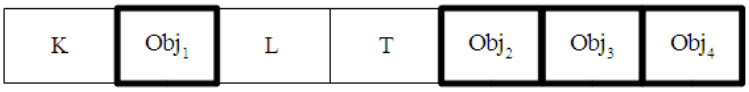

Рис 1. Пример общей структуры фразы

Для большинства ССТ, встречающихся в таксономических текстах, характерны три составные части [8]. Поэтому зададим структуру термина *obj* как вектор $obj =< o, P, obj' >$, где *obj* – корневой элемент; P – множество признаков корневого элемента; obj' – внутренний термин, зависимый от корневого элемента. Для наглядности введем 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». Выделим три основные части ССТ.

1. Корневой элемент o (ядро ССТ), который на семантическом уровне является классом терминов в эталонной таксономии, в который входит множество зависимых элементов. Элементы данного множества разделяются за счет использования в их описании различного рода признаков. На синтаксическом уровне это слово, которому подчиняется остальная часть описания термина. Это также означает, что остальная часть грамматически согласована с корневым элементом.

В используемом примере можно выделить два класса терминов:

- «повреждения», «изгибы», «изломы» относятся к одному классу понятий, объединяемых словом «повреждения»;
- «рельс», «подошва», «шейка» относятся к классу понятий, объединяемых словом «рельс».

2. Признаковая часть P, которая на семантическом уровне является суммой всех признаков, являющихся одним из способов определения занимаемого места среди множества элементов некоторого класса термина. На синтаксическом уровне, как правило, являются определениями (прилагательными, причастными оборотами, согласованными второстепенными предложениями). Кроме того, в признаковую часть могут входить ССТ, связанные с ядром предложно-падежной конструкцией. В используемом примере признаком является слово «поперечные», относящееся с корневому элементу «излом».

3. Субъект obj', который на семантическом уровне является значением, описываемым фразой, подчиненным ядру. С одной стороны, является частью родительского термина, а с другой – самостоятельным значимым термином. Имеет такую же структуру, как и весь ССТ, причем корневой элемент субъекта синтаксически согласован с корневым элементом данного термина. При этом каждый внутренний термин может относиться к различным классам предметной области.

Схематично структура ССТ показана на рис. 2.

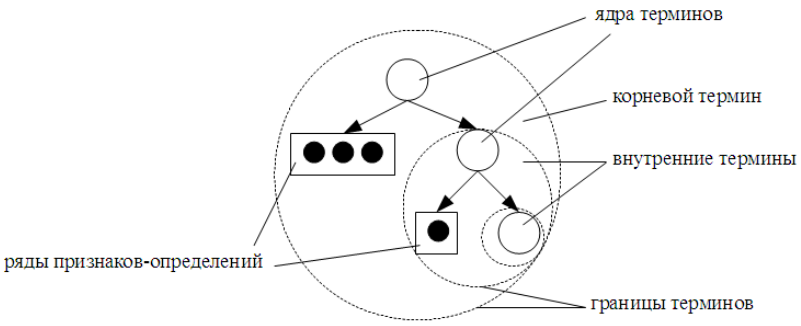

Рис. 2. Пример структуры сложного составного термина

Построение формального аппарата описания ССТ

Определим входное высказывание φ как упорядоченное множество слов $\varphi = (sw_1, ..., sw_n)$, где n – число слов в высказывании. У каждого слова есть собственные морфологические характеристики и синтаксическая роль относительно других слов в высказывании. Также каждое слово имеет собственное значение, которое может выражаться в тексте одним словом. Кроме этого, слово может быть частью словосочетания в высказывании, имеющего другое значение, не зависящее от данного слова.

На начальном этапе анализа необходимо определять для слов фразы соответствующие им части речи. Для этого зададим множество основных частей речи η, которое необходимо при анализе границ терминов и их связей: $\eta = \{\eta_{\text{сущ}}, \eta_{\text{прил}}, \eta_{\text{глаг}}, \eta_{\text{пред}}\}$ – и введем функцию $F_\eta(sw)$, которая возвращает элемент множества η для слова sw.

Зададим способы определения морфологических характеристик слов, описывающих ССТ. Существует два основных способа морфологического анализа: на основе словаря и на основе морфемного анализа [9]. Для достижения поставленных целей был использован подход на основе создания таблицы всех словоформ, т. к. он проще в реализации, а предметная область описывается конечным набором слов. Таким образом, словарь для слов, описывающих ССТ, представляется как множество словообразующих парадигм слова $D = \{dw_1, ..., dw_n\}$, где dw – словообразующая парадигма, определяемая как $dw = \{de_1, ..., de_m\}$, где $de_j^{dw_i}$ – один из видов представления слова в зависимости от рода, числа и падежа. Введем множества, определяющие набор морфологических характеристик для существительных и прилагательных: множество падежей $\sigma = \{\sigma_0, ..., \sigma_6\}$; множество рода $\tau = \{\tau_0, \tau_1, \tau_2\}$ и множество числа $\mu = \{\mu_0, \mu_1\}$. Также введем функции F_σ, F_τ, F_μ, позволяющие сопоставлять словоформу de с ее соответствующими характеристиками σ, τ, μ путем перебора множества словоформ. Кроме этого функция F_d сопоставляет слову sw в высказывании одну из словоформ морфологического словаря de. Таким образом, для каждого слова sw, описывающего термин, можно получить необходимые морфологические характеристики. Процесс получения информации о слове в процессе морфологического анализа приведен на рис. 3.

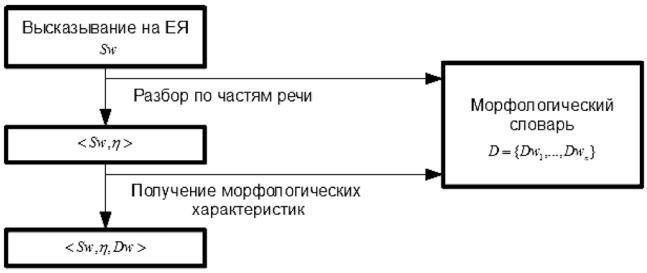

Рис. 3. Схема получения информации о слове в процессе морфологического анализа

Для того чтобы получить представление о структуре текста и входящих в него терминов, необходимо оперировать с синтаксическими характеристиками. Причем существует взаимосвязь между синтаксической ролью в предложении и местоположением в структуре ССТ. Поэтому введем предикат F_{lsync}, который определяет лингвистическую согласованность двух слов:

- $F_{lsync} : (sw_i, sw_j) \to \{0,1\}$. (1)

Для слов, описывающих ССТ, это означает следующее:

$$\sigma^{Sw_i} = \sigma^{Sw_j} \cup \tau^{Sw_i} = \tau^{Sw_j} \cup \mu^{Sw_i} = \mu^{Sw_j} \to F_{lsync}(sw_i, sw_j) = 1.$$

На основе предиката [10] можно задать предикат определения синтаксического подчинения, который позволит преобразовать упорядоченное множество слов в таксономическую структуру:

$$F_{sl} : (sw_i, sw_j) \to \{0,1\}.$$

Специальные предикаты позволяют делать предположения о семантической роли слова, опираясь на синтаксическую информацию. Однако особенности русского языка требуют нескольких критериев определения семантической роли, в том числе на основе заданных (эталонных) значений слова и словосочетания. Для критериев при необходимости можно определять степень значимости и порог реагирования. Введем множество критериев принадлежности Kr, элементами которого являются предикаты, определяющие принадлежность слова к определенной семантической роли:

$$Kr = \{kr_o^{syn}, kr_p^{syn}, kr_{sub}^{syn}, kr_o^{sem}, kr_p^{sem}, kr_{sub}^{sem}\},$$

где kr_o^{syn} – синтаксический (полученный на основе синтаксической информации) критерий ядра термина; kr_p^{syn} – синтаксический критерий признака; kr_{sub}^{syn} – синтаксический критерий субъекта; kr_o^{sem} – семантический (полученный на основе значения слова в эталоне) критерий ядра термина; kr_p^{sem} – семантический критерий признак; kr_{sub}^{sem} – семантический критерий субъекта.

В общем случае ядро является существительным и не имеет синтаксических зависимостей от других элементов термина. Внутри фразы не имеет зависимостей от

подлежащего и дополнения. Следовательно, можно обобщить критерий kr_o^{syn} для слова $sw_k \in \varphi$:

$$kr_o^{syn} = 1 \leftrightarrow (F_\sigma(sw_k) = \sigma_0) \cup (F_\eta(sw_k) = \eta_{\text{сущ}}) \cup (\bigcup_{i=1}^{n} (F_{sl}(sw_k, sw_i) \cup F_\sigma(sw_i) = \sigma_0) \neq 0).$$

Признаки не имеют зависимых слов, поэтому являются терминальными элементами. Поэтому критерий kr_p^{syn} для слова $sw_k \in \varphi$ задается как

$$kr_p^{syn} = 1 \leftrightarrow (F_\eta(sw_i) = \eta_{\text{прил}}) \cup (\bigcup_{j=1}^{n} F_{sl}(sw_i, sw_j) = 0).$$

Элемент термина – субъект s , в общем случае является дополнением в косвенном падеже, основным признаком этого элемента является отсутствие подчиненного слова. Поэтому критерий kr_{sub}^{syn} для слова $sw_k \in \varphi$ задается как

$$kr_{sub}^{syn} = 1 \leftrightarrow \bigcup_{j=1}^{n} (F_\sigma(Sw_k) \neq \sigma_0) \cup (F_\eta(Sw_k) = \eta_{\text{сущ}}) \cup F_{sl}(Sw_i, Sw_j) = 0.$$

Дополнение, которое имеет зависимость от ядра и вместе с тем имеет другое зависимое дополнение, образует новый термин obj' и становится его ядром. При этом как ядро o , так и простейший элемент s могут иметь неограниченное множество признаков P .

Полученные синтаксические критерии являются общими, их можно делить на составные высказывания и вводить систему их значимости. Таким образом, уже на этапе синтаксического анализа можно найти во фразе φ слова, относящиеся к множеству терминов obj и задать их структуру. На рис. 4 проиллюстрирована общая схема построения структуры сложного термина.

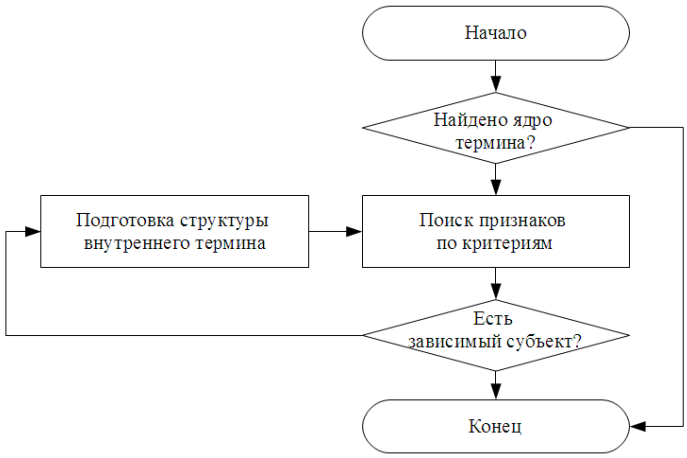

Рис. 4. Схема поиска ССТ в предложении

При этом данная функция возвращает одно наиболее вероятное значение. Реализация данной функции возможна, т. к. для составных частей терминов не так ярко выражена проблема омонимии. Причем множество *Sem* может описываться сложной системой значений, которая используется при оценке качества описанной таксономии, т. к. необходимо учитывать семантические связи между словами.

Для того чтобы оперировать с различными ССТ и его частями, объединим множество значений эталона в необходимую структуру. Так как структура эталонных знаний базируется на структуре субъективных знаний, изложенных в тексте, то обобщим рекурсивную структуру ССТ:

$$obj = < P_{obj}, o_{obj}, obj' > .$$

Если термин obj имеет внутренний термин obj' со схожей структурой с родительским термином, то имеет собственное ядро $o_{obj'}$, однако в косвенном падеже, т. к. оно подчинено родительскому ядру o_{obj}. Внутренний термин также может иметь свой внутренний термин obj''; если же его нет, то имеем ядро s, для которого нет подчиненных слов. Таким образом, получается система следующего вида:

$$obj' = \begin{cases} < P_{obj'}, o_{obj'}, obj'' >, \text{если} \quad obj'' \neq \varnothing; \\ < P_{obj'}, s_{obj'} >, \text{если} \quad obj'' = \varnothing,\ s_{obj} \neq \varnothing; \\ \varnothing, \text{если} \quad s_{obj} = \varnothing. \end{cases}$$

Исходя из структуры термина, зададим структуру хранения терминов в эталонной базе знаний. База знаний должна содержать термины, которые образуют таксономическую структуру. Каждый ССТ делится на элементы, являющиеся значениями, для которых задаются возможные текстовые выражения. Подобное деление позволяет задавать отдельное семантическое значение не только для слова, но и словосочетания. Это позволяет адекватно реагировать на различные именования одного и того же ССТ.

Введем понятие класса терминов Ω, в которых входят все термины с одинаковым ядром:

$$\Omega = \{obj_0, ..., obj_i, ..., obj_n \,|\, o_{obj_i} = o_{obj_j},\ i, j = 0..n\}.$$

Так как все термины класса имеют одинаковое ядро, то найденное во фразе ядро будет ассоциироваться с данным классом понятий. Следовательно, если ожидается соответствие между субъективными и эталонными знаниями, то в первую очередь в связи с ядром во фразе будут ожидаться элементы ядра в эталонной базе для данного класса.

Выделим семантические критерии, которые позволят определить местоположения термина во фразе, а также определить семантическую роль слова. Термин должен присутствовать в эталонной таксономии как класс понятий Ω, т. е. является ядром одной из семантик, причем конкретное семантическое значение определяется зависимыми элементами. Таким образом, семантический критерий для термина формулируется как

$$kr_o^{sem}(sem) = 1 \leftrightarrow sem \exists \Omega = \{obj \,|\, o^{obj} = sem\}.$$

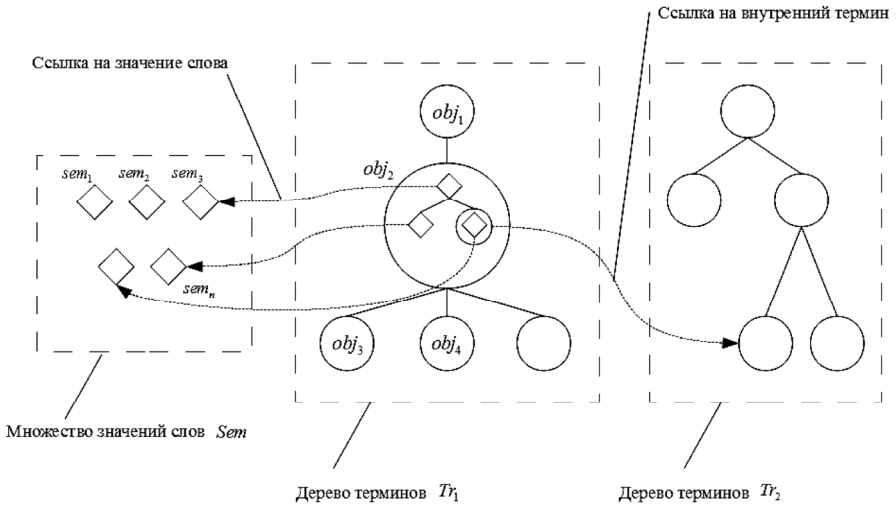

Рис. 5. Схематическое представление хранения значений в базе знаний

Если термин obj содержит в качестве субъекта внутренний термин obj', то эталонной базе знаний должны присутствовать описания обоих терминов, причем в описание общего термина obj включена ссылка на описание внутреннего термина obj' как его субъекта s^{obj}. При этом оба этих термина могут быть как из независимых деревьев, так и из одного дерева. Схематическое представление хранения значений в базе знаний приведено на рис. 5. Таким образом, семантический критерий для субъекта формулируется как

$$kr_{sub}^{sem}(sem) = 1 \leftrightarrow (sem \exists \Omega = \{obj \,|\, o^{obj} = sem\}) \cup (sem \exists \Omega^{sem^{obj}} = \{obj^{s} \,|\, s^{obj} = sem\}).$$

Для подтверждения того, что значение sem слова sw является признаком P^{obj} некоторого термина obj, нужно найти в эталонной базе знаний множество терминов Obj, к которым он принадлежит. Среди этого множества терминов предполагается такой, что его появление не нарушает последовательности описания таксономии:

$$kr_{p}^{sem}(sem) = 1 \leftrightarrow sem \exists Obj = \{obj \,|\, P^{obj} = sem\}.$$

Таким образом, введено множество критериев Kr, позволяющих определить семантическую роль слова, входящего в описание ССТ. Применяя критерии на этапах анализа текста, можно выделить из текста находящиеся в нем термины.

Заключение

Проанализирована структура высказываний, которые встречаются в описании таксономии в тексте на естественном языке. Полученные правила представления терминов заложены в основу структуры субъективной и эталонной баз знаний. Введены синтаксические и семантические критерии, позволяющие проводить автоматически разбор высказывания и принимать решения о семантической роли

того или иного слова высказывания. На основе анализа типичных высказываний для текста данного типа построен формальный аппарат, учитывающий основные особенности способов описания термина в высказываниях.

Библиографические ссылки

1. *Гаврилова Т. А., Хорошевский В. Ф.* Базы знаний интеллектуальных систем. – СПб. : Питер, 2000. – 384 с.

2. *Якимов В. Н., Мошков И. С.* Определение объектов и их характеристик в процессе обработки текстовой информации // Ресурсо- и энергосберегающие технологии и оборудование, экологически безопасные технологии : материалы девятой Междунар. науч.-техн. конф., Минск, 24–26 ноября 2010 г. : в 2 ч. – Минск : Белорус. гос. техн. ун-т, 2010. – Ч. 2. – С. 334–337.

3. *Гаврилова Т. А., Хорошевский В. Ф.* Базы знаний интеллектуальных систем.

4. *Знаков В. В.* Понимание в познании и общении. – Самара : СамГПУ, 2000. – 188 с.

5. *Гаврилова Т. А., Хорошевский В. Ф.* Базы знаний интеллектуальных систем.

6. Там же.

7. *Солсо Р. Л.* Когнитивная психология.

8. Там же.

9. *Леонтьева Н. Н.* Автоматическое понимание текстов: системы, модели, ресурсы. – М. : Академия, 2006. – 303 с.

10. *Гаврилова Т. А., Хорошевский В. Ф.* Базы знаний интеллектуальных систем.

***

V. N. Yakimov, Doctor of Technical Sciences, Professor, Samara State Medical University

I. S. Moshkov, Samara State Medical University

Structural Analysis of Compound Terms in Technical Documents

The features of natural language texts, describing a taxonomic structure are examined. The emphasis is made on classification of elements that are parts of compound terms found in the text. The article also defines membership criteria that help to classify a particular element of the compound term.

Keywords: natural language, text analysis, taxonomic structure

Получено: 15.11.11