УДК 519.767.6

СОПОСТАВЛЕНИЕ СИНТАКТИКО-ГРАММАТИЧЕСКОЙ И СЕМАНТИЧЕСКОЙ МОДЕЛЕЙ ТЕКСТА В ПРОЦЕССЕ АНАЛИЗА 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. Проведен системный анализ лексики текстов и определена связь основных лексических конструкций с их значением. На основе данного анализа формулируются правила сопоставления синтактико-грамматической и семантической моделей текста на естественном языке.

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

Информатизация науки и производства является объективным и неотъемлемым процессом современного постиндустриального общества. Поэтому актуальность разработки новых и совершенствования известных инструментов для извлечения информации постоянно растет. Одним из способов применения данных инструментов является оценка знаний, содержащихся в тексте [1, 2, 3], которая заключается в сравнении структуры знаний некоторого субъекта с эталоном и может использоваться как средство автоматической обработки результатов открытого тестирования.

Некоторые особенности текста на естественном языке (неполнота, избыточность, противоречивость) создают трудности в процессе создания инструмента для полноценного анализа текста [3, 4]. Таким образом, возникает потребность в 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 ряда 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 обеспечить структурирование и классификацию имеющихся знаний [5, 6, 7]. Следовательно, для решения задачи оценки знаний человека необходимо иметь систему распознавания терминов таксономии, которая описывается в документе на естественном языке.

В процессе достижения цели – автоматического сопоставления субъективных и эталонных знаний – решаются следующие задачи: анализ структурных особенностей текста таксономического типа; построение формального аппарата хранения знаний; определение критериев для сопоставления синтактико-грамматической и семантической моделей текста.

Для того чтобы сформулировать требования к формальному аппарату анализа, поделим высказывание на ЕЯ, описывающее таксономию, на отдельные части, и определим функции, которые они выполняют в тексте, а также возможные способы их нахождения. Ниже будем использовать высказывание $\varphi =< Obj, L, K, T >$, где Obj – множество сложных составных терминов (ССТ), L – связей между ними, K – критериев деления терминов, T – метаязыковых конструкций, описывающих качествен-

Илья Сергеевич Мошков – аспирант.

ные особенности таксономии. Для определенности в качестве примера будет использоваться следующее высказывание: «По химической классификации нефть делится на три основные группы: парафиновые нефти, нафтеновые нефти, ароматические нефти».

Для большинства ССТ, встречающихся в таксономических текстах, характерны три составные части [1, 6]. Поэтому зададим структуру термина *obj* как вектор $obj =< o, P, obj' >$, где o – 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) на семантическом уровне является классом терминов в эталонной таксономии, в который входит множество зависимых элементов. Под эталонной таксономией понимается экспертно заданное описание всех возможных классификаций предметной области. 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:

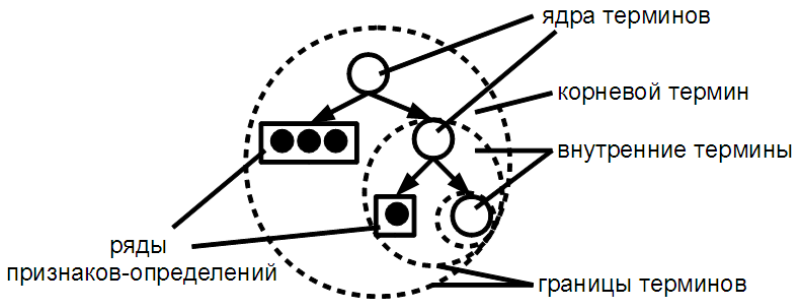
– «повреждения», «изгибы», «изломы» относятся к одному классу понятий, объединяемых словом «повреждения»;

– «рельс», «подошва», «шейка», относятся к классу понятий, объединяемых словом «рельс».

2. Признаковая часть P на семантическом уровне является суммой всех признаков, которые являются одним из способов определения занимаемого места среди множества элементов некоторого класса термина. На синтаксическом уровне они, 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*' на семантическом уровне является значением, описываемым фразой, подчиненным ядру. С одной стороны, он 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 (рис. 1).

Существует два основных способа морфологического анализа: на основе словаря и на основе морфемного анализа [3, 4]. Для достижения поставленных целей был использован подход на основе создания таблицы всех словоформ, так как он проще в реализации, а предметная область описывается конечным набором слов. Используем существующие методы морфологического и синтаксического анализа текста общего типа и применим их с учетом особенностей текста таксономического типа для извлечения его составных частей.



Р и с . 1 . Пример возможной структуры сложного составного термина

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}, определяющий лингвистическую согласованность текстового выражения слов sw_i и sw_j:

$$F_{lsync} : (sw_{i,} sw_j) \to \{0,1\}. \quad (1)$$

Для типов слов, обычно описывающих ССТ, характерно следующее:

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

где $\sigma^{Sw_i}, \tau^{Sw_i}, \mu^{Sw_i}$ означает падеж, род и число соответственно. На основе предиката (1) можно задать предикат определения синтаксического подчинения, который позволит преобразовать упорядоченное множество слов в таксономическую структуру:

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

Выделенные 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 признака, 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 $sw_k \in \varphi$:

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

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

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

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

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

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.

Выделяют несколько уровней значений набора слов – уровень слова, словосочетания, предложения и т. п. Поэтому эталонная система значений должна быть многоуровневой. Зададим систему значений на уровне слова и построим на этой системе значений множество необходимых для анализа уровней. Так как каждое слово sw является текстовым выражением определенного значения, то зададим систему, хранящую значения sem вводимого текста. Для сопоставления множества значений и множества их текстовых выражений введем функцию $F_{sem} : Sw \to Sem$ получения значения текстового представления. То есть если полностью задана система значений, то должно выполняться условие

$$\forall sw \exists sem, F_{sem}(sw) = sem.$$

При этом данная функция возвращает одно наиболее вероятное значение. Реализация данной функции возможна, так как 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, так как 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:

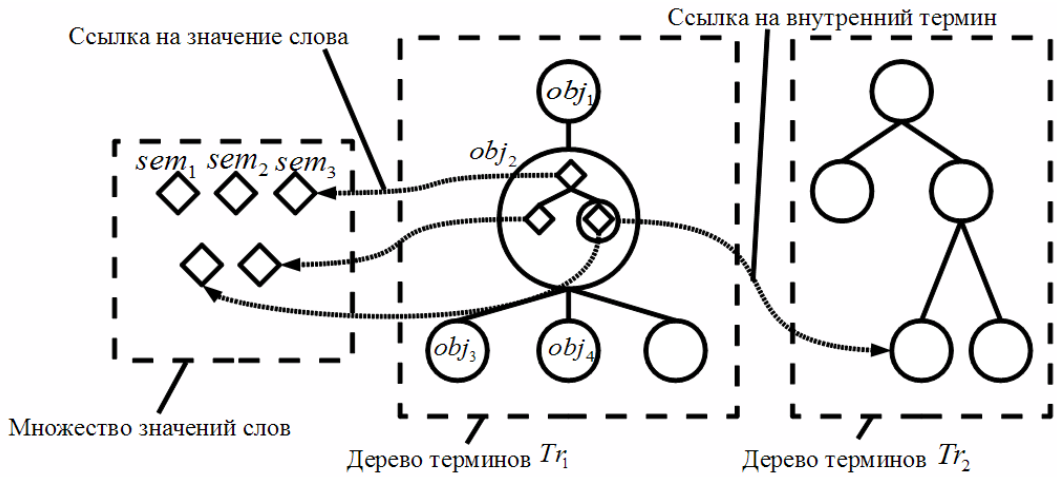
$$obj = < P_{obj}, o_{obj}, obj' > .$$

Если термин obj имеет внутренний термин obj' со схожей структурой с родительским термином, то имеет собственное ядро $o_{obj'}$, однако в косвенном падеже, так как 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 вида

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

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 для словосочетания. Это позволяет адекватно реагировать на различные именования одного и того же ССТ.



Р и с . 2 . Пример структуры эталонной базы знаний

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

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

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 (пример структуры приведен на рис. 2).

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

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

Если термин obj содержит в качестве субъекта внутренний термин obj', то в 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. Таким образом, семантический критерий для субъекта формулируется как

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

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:

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

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.

Если на основании критериев не удается подтвердить семантическое значение термина в анализируемом высказывании, то предполагается наличие допущенной ошибки в описании термина. Вариантов может быть несколько:

– распознанные слова термина не имеют синтаксической связки;

– слова термина имеют синтаксическую связь, но не имеют значения в данной предметной области на трех уровнях: ядра, одного или нескольких признаков, субъекта.

Таким образом, для синтаксически связанных элементов термина (P, o, s), не выполняется следующее условие:

$$E(P,o,s) = 1 \leftrightarrow P,o,s \exists Obj = \{obj \left| P \in obj \cup o \in obj \cup s \in obj\right\} \neq \varnothing.$$

Для определения степени ошибки введем предикаты ошибки использования того или иного элемента в термине. Предикаты допущенных ошибок в описании признаковой части термина E_p, при описании ядра термина E_o и описании субъекта термина E_s определяются как

$$E_p(P,o,s) = 1 \leftrightarrow E(P,o,s) = 1 \cup E(o,s) = 0,$$

$$E_s(P,o,s) = 1 \leftrightarrow E(P,o,s) = 1 \cup E(P,o) = 0,$$

$$E_o(P,o,s) = 1 \leftrightarrow E(P,o,s) = 1 \cup E(P,o) = 1.$$

При этом базовый алгоритм для определения степени ошибки в признаковой части опирается на отношение числа недопустимых элементов признаковой части ко всему количеству признаков. Так как ошибка может быть лишь в части термина, расчленение составных частей термина позволяет предположить подразумеваемое описание термина с учетом допущенной семантической ошибки, что позволяет ввести числовое значение степени допущенной ошибки.

В заключение можно отметить, что приведенная методика анализа производит поиск элементов таксономии и связей между ними в анализируемом тексте на основе набора синтаксических правил и эталонных знаний. Для поиска и проверки синтаксических и семантических конструкций используются особенности, характерные для текста таксономического типа. При этом есть возможность отбрасывать нераспознанную информацию, которая по ряду признаков не относится к описанию таксономии. Таким образом, предложенная методика анализа текста позволяет извле-

кать необходимую информацию как из текстов, непосредственно описывающих таксономию, так и из текстов, в которых присутствуют избыточные описания элементов таксономии или несущественная информация. Главным условием анализа является возможность построения на основе текста фрагмента таксономических знаний, сопоставимых с эталоном.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Никаев С.А.* Модели и информационная система для оценки профессиональных знаний специалистов промышленного производства / Автореф. дисс. … техн. наук Спец. 05.13.01. – Системный анализ, управление и обработка информации (промышленность). – Самара, 2004. – 24 с.
2. *Гаврилова Т.А*. Базы знаний интеллектуальных систем. – СПб.: Питер, 2000. – 384 с.
3. *Леонтьева Н.Н.* Автоматическое понимание текстов: системы, модели, ресурсы. – М.: Академия, 2006. – 303 с.
4. *Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л.* Лингвистический процессор для сложных информационных систем. – М.: Наука, 1992. – 256 с.
5. *Лурия А.Р.* Язык и сознание (Под ред. Е.Д. Хомской). – Ростов н/Д.: Феникс, 1998. – 416 с.
6. *Солсо Р.Л.* Когнитивная психология. – М.: Тривола, 1996. – 600 с.
7. *Знаков В.В.* Понимание в познании и общении. – Самара: СамГПУ, 2000. – 188 с.

Статья поступила в редакцию 15 июня 2011 г.

MATCHING SYNTACTICAL AND SEMANTIC TEXT MODELS IN NATURAL LANGUAGE TEXT ANALISYS

I.S. Moshkov

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

The article examines syntactical and semantic structure of taxonomy-based natural language texts. Was analysed texts vocabulary and connection between basic lexical structures and their meaning was determined. The analysis allows formalizing procedure of matching syntactical and semantic structure natural language texts.

Key words: *knowledge, natural language, taxonomy.*

Iliya S. Moshkov – Postgraduate student.