OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:(514+512)

ИНТЕГРИРОВАНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ АНАЛИЗА/СИНТЕЗА ИЗОБРАЖЕНИЙ И ТЕКСТА: КОНТУРЫ ПРОЕКТА INTEGRO

С.С. Курбатов (*kurbatow@yandex.ru*)
ОАО НИЦЭВТ, Москва, Россия

К.А. Найденова (*Naidenovaxen@gmail.com)*
Военно-медицинская академия, Санкт-Петербург, Россия

Г.К. Хахалин (*gkhakhalin@yandex.ru)*
независимый исследователь, Москва, Россия

В работе приводится описание языка представления онтологических знаний и структуры онтологии, кратко характеризуется экспериментальная прикладная область и дается описание интегрируемых интеллектуальных систем. По мере необходимости и возможности приводятся примеры, иллюстрирующие те или иные моменты изложения.

Ключевые слова: анализ и синтез изображений и естественного языка, интегрирование интеллектуальных систем, прикладная онтология, семантический гиперграф.

Введение

В работе представлен первый этап проекта INTEGRO (INTEGRating Ontology) объединения разномодальных интеллектуальных систем. Для этого этапа интеграции выбраны системы анализа/синтеза изображений, лингвистического анализа/синтеза текста и онтология, описывающая на языке представления знаний (общую для данных систем) информацию о мире. В дальнейшем предполагается в данную интегральную систему включить системы анализа и синтеза речи, поиска решений, порождение объяснений, индуктивного и аналитического (дедуктивного) обучения и другие.

Интерфейсом для интегрируемых систем служит общая для них (прикладная) онтология, описанная на языке семантического гиперграфа. В качестве экспериментальной прикладной области выбрана среда с условным названием «Планиметрия» - область, богатая как на изображения (плоские фигуры и их комбинации), так и на текстовые описания объектов данной реальной среды (формулировки на естественном языке планиметрических построений). По мере расширения интегральной системы будут привлекаться и другие предметные области.

Попытки объединения в некоторых (ограниченных) комбинациях систем, например, синтез изображений по описаниям на естественном языке, представлены в [Tandareanu N. et al., 2003], [Wang J. et al., 2009]. Или в системах машинного перевода – анализ и синтез предложений естественного языка (переводчики ПроМТ, Google и др.). При этом большинство современных лингвистических анализаторов/синтезаторов в системах перевода не «работают» с «сильным» семантическим результатом, «поворот» от анализа к синтезу происходит на уровне синтаксических структур или на уровне так называемой поверхностной семантики, что, естественно, отражается на качестве перевода.

Представляемые в проекте комплексные системы могут найти применение в обучающих системах по различным дисциплинам с привлечением графической и естественно языковой

информации, в системе сурдоперевода (текст в жест), в системах поиска изображений по текстовому описанию и в других областях. Но наиболее важный (на наш взгляд) результат интеграция полного спектра разномодальных систем дала бы в робототехнике. Нельзя сбрасывать со счетов и чисто научные цели, поскольку интегрирование интеллектуальных систем ставит много неизученных вопросов и, в частности, взаимодействие интегрируемых систем, расширенная постановка задач (совместного) понимания текста и изображений, постановка задачи «комплексного» (индуктивного и дедуктивного) обучения и т.д. Эти задачи для области искусственного интеллекта являются новыми и требуют подчас нестандартных решений и «незамыленного» взгляда на проблематику.

В докладе приводится описание языка представления онтологических знаний и структуры онтологии, кратко характеризуется экспериментальная прикладная область и дается описание интегрируемых интеллектуальных систем. По мере необходимости и возможности приводятся примеры, иллюстрирующие те или иные моменты изложения.

1. Прикладная онтология

Под онтологией подразумеваем концептуальную (в широком смысле слова) «модель мира». Прикладные онтологии описывают концепты, которые зависят как от онтологии задач, так и от онтологии предметной области [Гаврилова, 2006]. Онтологический инжиниринг при этом подразумевает: определение классов понятий; наведение таксономии на классах; разработку структур понятий и ситуаций; определение свойств понятий и значений этих свойств; процедуры вывода на онтологии и преобразования описаний в модели.

Если в качестве единого интерфейса интегрируемых систем берется онтология, то возникает два вопроса. Первый вопрос касается выбора языка представления онтологических знаний, чтобы он в дальнейшем позволил «погружать» в него требуемые расширения, например, нечеткие знания, познавательные процедуры и т.д. Второй вопрос относится к разработке прикладной онтологии общей для систем, способной на концептуальном уровне интегрировать разномодальные входы для синтезаторов и выходы для анализаторов интегральной системы. При этом, конечно, предполагается, что «внутри» каждой системы могут быть свои языки представления и свои базы «внутренних» знаний. Например, для зрительной системы анализа изображений может существовать своя база для выделения «непроизводных» объектов; для системы анализа текста – свои для морфологии и синтаксиса естественного языка (ЕЯ).

1.1. Язык представления знаний

Для разных предметных областей и для разных задач существует спектр языков (моделей) представления знаний. Обзоры некоторых из них даны в [Башмаков и др., 2006]. На наш взгляд наиболее адекватным языком представления концептуальных знаний для разномодальной информации является язык гиперграфов в качестве расширения семантических сетей, где естественным образом представляются *n*-арные отношения, которые позволяют задавать не только атрибуты объектов, но и представлять структурные, «целостные» описания объектов с взаимосвязями своих компонентов.

Известно [Зыков, 1974], [Визинг, 2007], что гиперграф H (V, E) определяется парой, где V – множество вершин $V = \{v_i\}$, $i \in I = \{1, 2,\ldots, n\}$, а E – множество ребер $E = \{e_j\}$, $j \in J = \{1, 2,..., m\}$; каждое ребро представляет собой подмножество V. Вершина v и ребро e называются *инцидентными*, если $v \in e$. Для $v \in V$ через $d(v)$ обозначается число ребер, инцидентных вершине v; $d(v)$ называется *степенью вершины* v. *Степень ребра* e – число вершин инцидентных этому ребру, - обозначается через $r(e)$.

Для гиперграфа можно ввести понятие инцидентора. Пусть e — ребро, которому инцидентна вершина v. Тогда пара (v,e) называется *инцидентором* этого ребра, *примыкающим* к вершине v (или, *при* вершине v). Два различных инцидентора одного и того же ребра называются *сопряженными*. Два различных инцидентора, примыкающих к одной и той же вершине, называются *смежными*.

Гиперграф H является *r-однородным*, если все его ребра имеют одинаковую степень r. Если каждое его ребро имеет степень равную 2, то гиперграф является графом.

Гиперграф (V', E') называется *подгиперграфом* (V, E), если $V' \subseteq \boldsymbol{V}$, $E' \subseteq E$ и вершина $v \in V'$ и ребро $e \in E'$ инцидентны в (V', E') тогда, и только тогда, когда они инцидентны в (V, E).

Гиперграфы бывают *ориентированные* и *неориентированные*. Рёбра неориентированного гиперграфа называются *звеньями*. В случае ориентированного гиперграфа (оргиперграфа) ребро $e \in E$ называется *гипердугой* (для краткости *дугой*) и представляется упорядоченной парой (h, T), где $h \in V$, $T \subseteq V \backslash \{h\}$, $T \neq \varnothing$. При этом вершина h называется *началом* дуги e, а каждая вершина из T — *конечной вершиной* дуги e. Говорят, что дуга e *исходит* из вершины h и *заходит* в каждую из вершин множества T. Если в гиперграфе присутствуют звенья и гипердуги, то гиперграф называется *смешанным*. На рисунке 1 показан общий вид гиперграфа $H(V, E)$, где $V = \{V_1, V_2, \ldots, V_6\}$; $E = \{E_1=\{V_1\}, E_2=\{V_1,V_3\}, E_3=\{V_1,V_2,V_3\}, E_4=\{V_2,V_4\}, E_5=\{V_2,V_4\}, E_6=\{V_3,V_4,V_5\}, E_7=\varnothing\}$.

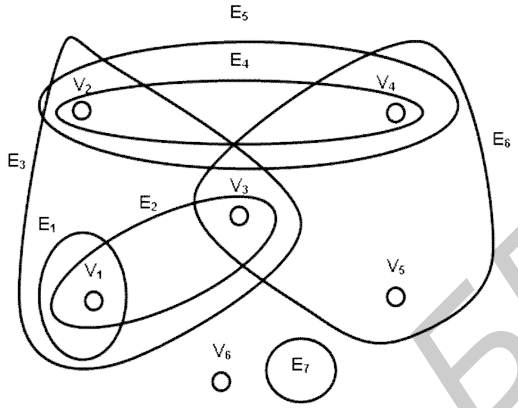

Рисунок 1 – Общий вид гиперграфа

Если элементам гиперграфа приписать цепочки символов из некоторого множества, то он будет гиперграфом с помеченными (раскрашенными) вершинами и ребрами. Цепочки символов – это имена понятий и отношений онтологии, представленной гиперграфом. Такой гиперграф будем называть *семантическим гиперграфом*. Пример описания понятия *Треугольник* со структурой в виде семантического гиперграфа представлен на рисунке 2.

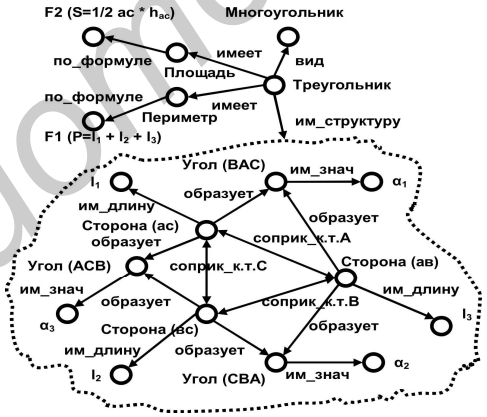

Рисунок 2 – Структурное описание понятия *Треугольник*

1.2. Экспериментальная прикладная область

Среда «Планиметрия» выбрана из-за возможности отрабатывать на ней зрительное восприятие, синтез зрительных объектов, анализ и синтез текстов. В дальнейшем можно наращивать интегральную систему блоками планирования решения задач, «ответов на вопросы», «объяснения», привлекать познавательные процедуры (индукцию, аналогию и др.), а также ставить задачу обучения и самообучения как развитие известных методов, приближаясь к «человеческому» обучению.

Объектами среды являются «чисто» планиметрические фигуры (отрезок прямой, различные виды треугольников, прямоугольник, ромб, трапеция, эллипс, круг, угол, медиана и т.д.) и плоские «детские» картинки (кораблик, домик, паровозик, цветочек, ромашка, рожица,

человечек и т.п.). Часть изображения может представлять собой фрагменты ЕЯ-текста и естественно-языковые надписи (обозначение сторон объекта, наименование объекта и т.п.).

Ситуации «компонуются» из некоторого множества взаимосвязанных объектов, которые могут быть представлены на изображении или выражены ЕЯ-текстом.

Отношения в среде носят разнообразный характер. Они включает различные группы: контактные (*соприкасается*, *пересекает*), пространственные (*правее*, *справа внизу*, *следует за*), метрические (*имеет длину*, *имеет толщину*) и другие.

Естественный язык (ЕЯ) для данной области включает в себя не только сравнительно четкие формулировки планиметрических задач (хотя и здесь есть много неопределенностей), но и описания на ЕЯ рисунков типа *кораблик*, при которых «вариабельность и произвол» выражения достаточно широки. Лексика этого подмножества языка ограничена предметной областью, смысл слов более или менее однозначен, отсутствуют метафоры, ассоциации и т. п. Тем не менее, в нем используется широкий спектр свойств ЕЯ и конструкций предложений, включая омонимию, полисемию, осложненные и сложные предложения, неполные конструкции (эллиптические и анафорические предложения), ошибки, специальные вкрапления (формулы, геометрические и математические знаки) и т.д. Достаточно часто естественно языковые выражения для анализа состоят не из одного изолированного предложения, а из нескольких, составляя фрагменты текста. Ниже приведены примеры разнообразных геометрических формулировок с различными сложностями и неопределенностями:

Периметр прямоугольного треугольника равен 132, а сумма квадратов сторон треугольника – 6050. Найти стороны. (Здесь тире стоит вместо слова равна, а в предложении *Найти стороны* пропущено слово *треугольника*).

Для каждого угла треугольника существует ровно два смежных с ним внешних угла этого треугольника (предикат *смежен* не имеет смысла для треугольника).

В треугольник, периметр которого равен 18 см, вписана окружность, в которой проведена касательная параллельно основанию треугольника. Отрезок касательной, заключенный внутри треугольника, равен 2 см. Вычислите основание треугольника. (вульгаризм «*вычислите основание*», образованный за счет опущенного слова: *Вычислите длину основания треугольника*).

Докажите что квадрат площади четырехугольника описанного около окружности равен $abcd\sin^2(\beta+\delta)/2$.

Данный треугольник разделите на две равновеликие части прямой, параллельной одной из сторон. (равновеликий = равный по площади).

Докажите, что основание высоты прямоугольного треугольника делит его гипотенузу на отрезки, пропорциональные квадратам катетов.

Дано: точка M лежит внутри $\angle BOA$, $\angle BOM = 40°$, $\angle MOA = 25°$, $MB \perp OB$, $MA \perp OA$. *Найти углы треугольника MBA.*

Точка M(x,y) принадлежит окружности, если ее координаты x и y удовлетворяют уравнению $(x-x_0)^2+(y-y_0)^2=R^2$.

1.3. Структура прикладной онтологии

Прикладная онтология «Планиметрия» разрабатывается на основе общих принципов построения онтологий. Формализм семантических гиперграфов позволят определить онтологию в виде:

O (H, I, P), где

H (X, R) – семантический гиперграф для предметной области, X – множество понятий проблемной среды со своими в общем случае структурами и свойствами (множество вершин гиперграфа), R – множество отношений между понятиями (дуги и ребра гиперграфа), I – множество имен понятий и отношений в данной предметной области, а P – множество процедур вывода на онтологии.

Множество понятий проблемной среды разделяется на несколько подмножеств:

$X = \{X_1,.., X_k\}$, где

X_1 – это класс подклассов фрагмента проблемной среды (например, класс *Плоская Фигура*). X_2 – это подклассы «структурных» понятий проблемной среды (*Треугольник*, *Трапеция*, *Кольцо*

и др.), X_3 – классы «составляющих» структурные понятия (*Сторона*, *Основание*, *Катет* и др.), X_4 – свойства понятий (*Длина, Высота, Периметр*, *Площадь* и т.д.), X_5 – значение свойств (значение угла α_i, значение длины стороны l_j) и др.

На семантическом гиперграфе можно представлять и результаты арифметических, логических и других операций, поэтому в онтологии можно вводить соответствующие вершины (например, *Разность*, *Сумма*, а также формулы подсчета периметра, площади и т.п.). Множество понятий *X* является открытым – его можно расширять по мере необходимости.

Отношения между классами, подклассами и надклассами понятий организуются в виде *таксономии* или *таксономической иерархии*. Для представления таксономии используется отношение *является_видом* (*AKindOf*).

При разработке и при использовании любой онтологии необходимо определить перечень используемых отношений. На сегодняшний день нет общепринятого полного перечня отношений за исключением десятка общезначимых отношений (например, *A_Kind_Of*, *part_of*, *have_value*, *have_structure* и др.). Тем не менее, всю совокупность отношений в онтологии стоит разделить на несколько подмножеств:

$R = \{R_1,\ldots, R_m\}$, где

R_1 – *общезначимые* отношения (см. выше), R_2 – *арифметические*, R_3 – *логические* (И, ИЛИ, НЕ и др.) и т.д. и R_k – *предметные* для данного приложения отношения. Общезначимые отношения носят в основном декларативный характер: они служат для нахождения путей поиска требуемых знаний. Арифметические, логические, функциональные и предметные отношения могут интерпретироваться (в зависимости от системы, входящей в комплекс) как декларативно, так и процедурно. Процедурная интерпретация означает, что с именем данного отношения связана соответствующая процедура, с помощью которой осуществляются некоторые действия в реальной или «виртуальной» среде (это аналогично присоединенным процедурам в теории фреймов). Например, отношение *соприкасается_в_концевой_точке* интерпретируется зрительной системой как присоединенная процедура, позволяющая на изображении найти пересечение. А вот для системы синтеза текста отношение *делит_пополам* будет интерпретироваться как декларативное словосочетание *делит пополам*. Отметим, что множество *R* является также открытым множеством.

Имена понятий и отношений онтологии (множество *I*) выбираются на основе терминологии, соответствующей естественно-языковым реалиям в данной области, и из-за необходимости упрощения процесса разработки онтологии для эксперта и инженера по знаниям. Но здесь процесс стандартизации далек от завершения, поэтому каждый разработчик предметной онтологии по возможности использует общепринятую терминологию (в основном для имен понятий) и собственные предпочтения – для имен отношений. В планиметрии существует устоявшаяся терминология для обозначения понятий (еще со времен Древней Греции). И понятия обозначаются цепочками символов очень похожими на слова естественного языка, например, либо английского, либо русского. С наименованием отношений – вопрос сложнее, поэтому и разнобоя больше.

Множество процедур вывода *P* на онтологии подробно описано в [Хахалин, 2009].

Фрагмент онтологии *Плоская фигура* с таксономией, с (незаполненными) структурами фигур и со свойствами концептов на языке гиперграфов приведен на рисунке 3.

Описание ситуаций в онтологии чаще всего не имеют статуса понятия, поскольку они носят временный характер. Иногда ситуации (скажем, в зависимости от частоты их появления) могут оформляться в онтологии в качестве некоторого сложного понятия. Примерами ситуаций могут быть: сторона прямоугольника совпадает с катетом треугольника; рядом с домиком справа располагается елка; окружность, вписанная в треугольник и т.д.

Экземпляр объекта в онтологии в общем случае представляет собой структуру с полностью или частично означенными параметрами. На рисунке 4 приведен экземпляр структуры понятия *Равнобочная трапеция* с означенными длинами оснований и боковых сторон. Означенные параметры в графе выделены черными кружками.

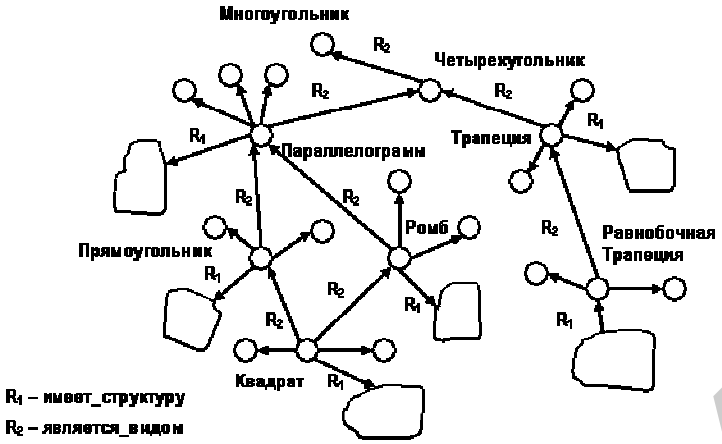

Рисунок 3 - Фрагмент онтологии *Плоская фигура* с таксономией, структурами фигур и со свойствами концептов

Рисунок 4 – Частично означенный экземпляр понятия *Равнобочная трапеция 1*

2. Система анализа изображений

Под задачей анализа изображения будем понимать описание геометрических объектов и их расположение на изображении в терминах и в структурах языка прикладной онтологии, а не просто отнесение объектов на изображении к известным классам. Результатом анализа должно быть описание ситуации, состоящей из экземпляров распознанных классов объектов с их означенными характеристиками и отношениями между ними.

Интегрирование адекватно предполагает гибридизацию систем, если под гибридизацией понимать процесс создания сложной системы, в которой интегрируются знания и традиционная обработка. Исходя из этого, систему анализа изображений рассматриваем как гибридную систему, состоящую из системы распознавания (в классическом смысле) для «непроизводных» объектов и систему концептуального анализа объектов, определенных на «непроизводных» объектах и на объектах более сложного характера, представленных в онтологии.

Граница между «непроизводными» и сложными объектами достаточно условна. Отрезок прямой и дуга - «непроизводные» объекты, а трапеция и другие фигуры – сложные объекты. Или отрезок прямой, дуга, трапеция, прямоугольник, ромб и т.п. - «непроизводные» объекты, а сложными объектами будут: кораблик, паровозик, кольцо, окружность, вписанная в треугольник и т.д. Поэтому в онтологии дублируются описания «непроизводных» объектов в концептуальном виде, а не в виде кортежа признаков. То есть в гибридной системе

концептуальный уровень анализа изображений дублирует систему распознавания образов (для надежности, для выявления неопределенностей, для коррекции результатов предыдущего этапа), но использует преимущества системы распознавания образов. А к преимуществам распознавания образов относятся относительная эффективность (скорость), «прозрачная» постановка задачи, наработанный потенциал и имеющие признание методы обучения.

2.1. Распознавание «непроизводных» объектов

Процесс извлечения знаний из данных можно представить как пошаговый и многоуровневый процесс, при котором происходит преобразование концептов более низкого уровня к концептам более высокого уровня, причем это преобразование происходит на основе одних и тех же принципов не зависимо от уровня генерализации и природы данных.

Концепты или паттерны более низкого уровня с их выделенными признаками служат исходными данными для выделения концептов или паттернов более высокого уровня.

Есть принципиально два пути выделения концептов следующего более высокого уровня иерархии: использование программных модулей, воплощающих известные математические методы и знания специалистов об инвариантных свойствах выделяемых концептов (выделение однородных областей, выделение границ областей, контуров объектов, кластеризация и т.п.) и использование индуктивных методов обучения концептам по примерам. Первый путь применяется чаще всего к выявлению объектов самого низкого уровня. Методы индуктивного обучения применимы в том случае, когда удается концепты более низкого уровня описать с помощью заданного набора признаков (атрибутов). К этим методам относится, например, машинное обучение концептам по примерам.

Метод обучения подразумевает большую активность эксперта в управлении процессом извлечения объектов из изображений: от задания обучающей выборки до использования правил, моделирующих рассуждения специалистов. Эти правила включают [Naidenova, 2004]: классификацию или разбиение объектов на непересекающиеся классы по значениям некоторого признака, сравнение, вычисление разницы в значениях числовых признаков, специализацию или добавление к описанию значений новых признаков, диагностику или выделение значений признаков, различающих заданные объекты, генерализацию или выделение общего признака, кластеризацию и т.п. В качестве очень простого примера опишем формирование концепта «прямоугольник» при анализе выделенных линий на изображении.

В результате линеаментного анализа изображения (= анализа линий на изображении) имеется коллекция линий, описание которых содержит множество координат точек, лежащих на линии, и направление линии или угол между линией и заданной координатной осью. Затем формируются концепты первого уровня иерархии: «линии, имеющие одно и то же направление или параллельные линии», «две пересекающиеся линии» и «две перпендикулярные линии». Для формирования первых двух концептов применяется операция выделения общего признака линий (генерализация), для формирования третьего концепта необходима операция сравнения и вычисление разности между направлениями линий. Одновременно с описанием концепта (правилом его определения) формируется и процедура для его выделения.

Концепт следующего уровня иерархии – «прямоугольник» строится с помощью концептов предыдущего уровня. Для получения этого концепта и построения процедуры поиска всех прямоугольников на изображении используются следующие операции: классификация линий по признаку «все линии одного направления»; выделение двух классов линий, с разницей направлений, равной 90°; выделение для линии одного класса множества INTERSET всех линий другого класса, которые её пересекают; нахождение двух параллельных линий a, b одного класса, для которых множества INTERSET(a), INTERSET(b) пересекаются. Результатом является описание целевого концепта и последовательность операций, по выделению этого концепта на изображении. Пользователь может порождать различные, но эквивалентные процедуры получения одного и того же концепта. Иерархическая структура концептов образует в конечном итоге объектно-ориентированную систему знаний, которую можно представить и как сетевую и как реляционную структуру.

Реализация процесса иерархического синтеза сложных геологических структур из элементарных структур с помощью экспертных продукционных правил интерпретации

изображений описывается в работе [Denisov at al., 1991]. Ряд экспертных правил можно получать с помощью операций машинного обучения.

2.2. Программа MyScript Notes

Программа MyScript Notes фирмы VisionObjects [MyScript Notes, 2007] преобразует и по мере возможности распознает рукописный текст, фигуры и таблицы. Она имеет несколько режимов работы: «Форматированный текст» и «Графика и текст». В последнем режиме она преобразует в различных комбинациях текст, фигуры и рисунки произвольной формы.

Программа распознает только рисованные геометрические фигуры. Они рисуются с помощью пера в одном окне, а распознанные фигуры перерисовываются и отображаются в другом окне программы с большей четкостью и плавностью их начертания. В режиме «Графика и текст» обрабатываются фигуры: окружность, эллипс, прямоугольник, ромб, треугольник, дуга, отрезок прямой, прямые и изогнутые стрелки, табличные структуры и рисунки произвольной формы (но нет трапеции!?). Если программа Notes не в состоянии определить фигуру, она пытается преобразовать ее в рисунок произвольной формы.

Полученный в результате преобразования файл содержит информацию в цифровом виде (на своем языке представления данных) и доступен для использования (экспорта). Кроме имени класса распознанной фигуры программа фиксирует конкретные значения соответствующих параметров фигуры в зависимости от ее класса (координаты вершин, центр и радиус окружности и т.п.).

Общая структура описания фигур в выходном файле выглядит следующим образом:

<имя фигуры> <вероятность распознавания> <параметр фигуры 1> … <параметр фигуры n>

Элементы описания отделяются друг от друга пробелами. Ниже приведены описания некоторых фигур с примерами:

круг <вероятность распознавания> <координаты центра окружности: X Y> <радиус круга>

circle 0.547534 721.655 536.568 174.44

прямоугольник <вероятность распознавания> <точка вершины1: X1 Y1> <точка вершины2: X2 Y2> <точка вершины3: X3 Y3> <точка вершины4: X4 Y4>

rectangle 0.417602 901.834 659.895 573.333 691.613 545.872 407.208 874.373 375.49

параллелограмм <вероятность распознавания> <точка вершины1: X1 Y1> <точка вершины2: X2 Y2> <точка вершины3: X3 Y3> <точка вершины4: X4 Y4>

parallelogram 0.417359 904.744 660.886 576.047 690.166 542.875 406.359 871.572 377.079

треугольник <вероятность распознавания> <точка вершины1: X1 Y1> <точка вершины2: X2 Y2> <точка вершины3: X3 Y3>

triangle 0.186418 569.912 694.998 1105.94 633.292 543.062 315.433

отрезок прямой <вероятность распознавания> <концевая точка1: X1 Y1> < концевая точка2: X2 Y2>

line 0.631655 635.269 178.377 689.992 88.1751

Для каждого изображения в выходном файле задается информация о возможных кандидатах на распознавание с соответствующими вероятностями. Например, для изображения *кораблик с флажком* на рисунке 5 распределение кандидатов выглядит следующим образом:

parallelogram 0.841063 528.52 912.446 747.868 908.344 840.673 554.967 621.324 559.07

drawing 0.5

triangle 0.31167 497.79 1142.93 906.228 557.91 608.597 556.68

ellipse 0.290373 685.072 726.389 211.211 122.372 -1.14976

arc 0.290167 685.072 726.389 211.211 122.372 -1.14976 0 6.28319

line 0.731522 353.853 1307.62 956.88 1293.76

drawing 0.5

arc 0.100776 636.153 1305.02 253.049 6.93823 -0.0208331 2.56482 6.28319

arrow 0.58867 728.269 1086 728.952 513.004 724.757 405.722 719.983 536.781 838.554 471.303 0

drawing 0.5

arc 0.155304 771.232 752.132 371.109 59.9316 1.55854 0.503146 3.89018

line 0.834599 261.09 1084.42 1079.87 1067.7

drawing 0.5

arc 0.125702 660.368 1072.95 392.703 6.51846 -0.00942964 -1.11674 3.12932
drawing 0.5
line 0.436261 267.309 1072.13 362.242 1311.92
arc 0.201896 307.866 1192.53 118.11 9.32409 1.19305 3.09251 6.28319
drawing 0.5
line 0.400571 724.831 518.537 838.117 480.311
arc 0.121137 784.761 499.124 52.1676 3.29824 -0.349319 3.12932 7.20357
drawing 0.5
line 0.424352 957.14 1296.5 1085.65 1070.8
arc 0.159768 1017.81 1180.51 114.253 7.37352 -1.03893 -0.638136 3.14159

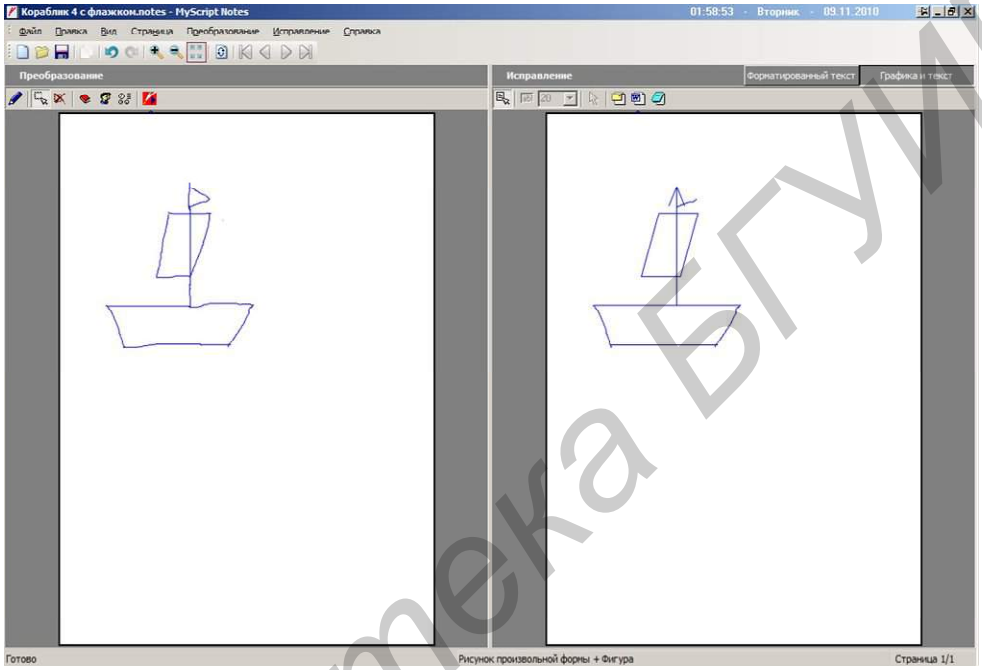

Рисунок 5 – Пример распознавания изображения *кораблик с флагом*

Результаты распознавания – это имена классов фигур (выделенных полужирным шрифтом) с соответствующими параметрами: параллелограмм, отрезок прямой (2 шт.), стрелка и произвольная кривая (4 шт.).

2.3. Концептуальный целенаправленный анализ изображений

Какая бы не была выбрана система распознавания образов, полностью на нее возложить ответственность за решение задачи концептуального анализа изображения нельзя, поскольку результат распознавания не дает возможность избавиться от недостатков системы распознавания. К недостаткам можно отнести: переход к пространству признаков «сразу» приводит к потере информации; добавление нового класса ведет к переобучению системы; предварительная обработка осуществляется над всем изображением и т.п. Поэтому в качестве охватывающей «надстройки» над системой распознавания образов стоит система концептуального целенаправленного анализа.

В основе организации процесса целенаправленного анализа изображений лежит использование прикладной онтологии структурных описаний, адекватных анализируемым изображениям не только относительно задачи анализа, но и задачи синтеза [Байков и др., 1980].

Структурность описания объекта подразумевает, что рассматриваемый объект как целое состоит из частей, связанных между собой определенными пространственными отношениями.

Адекватность этих описаний процессу распознавания удовлетворяет принципу двойного соответствия, которое гласит, что описание должно быть необходимым и достаточным как для распознавания объектов изображений, так и для возможности построения по этому описанию множества различных экземпляров объекта.

Такие описания обладают свойством полноты, т. е. в них задается полная информация о классе объектов (свойство независимости описаний), а не только та, которая отличает представителей данного класса от представителей других классов. Независимость описаний определяет открытый характер онтологии: в нее можно добавлять описания новых объектов, в ней можно корректировать уже представленные описания, проводить всевозможные связи между фрагментами модели и т.д. без нарушения немодифицируемых частей онтологии.

С каждым исходным элементом ассоциируется процедура выделения этих элементов непосредственно на изображении. Алфавит исходных «непроизводных» и «производных» элементов включает понятия *точка*, *отрезок*, *треугольник*, *трапеция*, *круг*, *эллипс*, *дуга*, *кораблик*, *домик*, и т.п. Алфавиты исходных элементов и отношений открыты и могут быть расширены (модифицированы, заменены) без перестройки системы анализа.

Структура онтологии реализует в процессе распознавания целенаправленное взаимодействие следующих стратегий: от результатов распознавания образов к модели, «сверху-вниз» (от модели к изображению) и «снизу-вверх» (от изображения к модели). То есть система концептуального анализа может работать как независимо от системы распознавания образов, так и на основе ее результатов. А в общем случае все время происходит процесс взаимодействия этих систем.

Процесс контекстного анализа изображения носит итеративный характер:

- выдвижение гипотез о присутствии в определенном месте изображения представителя того или иного класса распознаваемых объектов;
- проверка выдвинутых гипотез;
- принятие решений (текущих и окончательных) о справедливости той или иной гипотезы.

Выдвижение и проверка гипотез управляется контекстом, образуемым текущим результатом, информацией из прикладной онтологии и априорной информацией.

Проверка гипотез сводится к целенаправленному поиску непосредственно на входном изображении совокупности элементов, удовлетворяющих структурному описанию данного класса анализируемых объектов.

Сегментация изображения на отдельные фрагменты, а также интерпретация этих элементов осуществляются непосредственно в ходе анализа и проводятся в рамках проверяемой гипотезы. Предварительной сегментации и описания в признаках не требуются. Единственной необходимой операцией на изображении является фильтрация (избавление от шума).

Решение о справедливости гипотезы может быть пересмотрено на последующих этапах процесса анализа, если этого потребует контекст более высокого уровня, чем объект, связанный с данной гипотезой.

Результатом работы всей системы анализа изображения является описание зрительной ситуации на концептуальном языке онтологии. Уровень описания определяется самим изображением и онтологией на данный момент времени.

Проиллюстрируем этапы контекстного анализа на примере распознавания объекта *кораблик с флагом* (см. рисунок 5). Пусть в онтологии присутствует структура понятия *трапеция*, *кораблик* и *кораблик с флажком*. Последняя структура в виде текста на ЕЯ могла бы выглядеть следующим образом:

остов кораблика (= трапеция), *к большему основанию которой примерно в центре касается мачта* (= отрезок прямой) *и мачта перпендикулярна основанию. На мачту «нанижен» парус* (= параллелограмм). *К концу мачты присоединен флажок* (= треугольник).

Тогда на основе результатов распознавания по связям типа *входит_в_структуру* (от понятий *параллелограмм* и *отрезок*) в качестве структуры-гипотезы будет выбрана структура *кораблик*. Сопоставление структуры *кораблика* с результатами распознавания даст положительный результат (четыре отрезка с учетом координат концевых точек составляют *трапецию*, *отрезок* соответствует *мачте*, а распознанный *параллелограмм – парусу*). Но поскольку на изображении остались еще «не привязанные» элементы (*стрелка – arrow* и маленький «хвостик» - *отрезок прямой*, см. изображение), то система ищет контекст более высокого уровня, куда входит понятие *кораблик*. Таким понятием будет понятие *кораблик с флажком*. Сопоставление этого понятия с результатами распознавания не даст положительного результата. Это будет сигналом к тому, что либо онтология неполна, либо часть изображения распознано неправильно. Приняв за основу последнее, система передает управление

программе, которая будет «оставшиеся» части объекта искать на изображении, уже зная, где примерно искать и что искать (отметим заодно, что такая стратегия одновременно хорошо решает задачу сегментации изображения).

Область поиска определяется координатами объекта *стрелка* и «хвостиком» (см. эти параметры перед рисунком 5). А объектом поиска будет *флажок* (= треугольник). При такой реализации поиска объектов на изображении верхний уровень анализа целенаправленно может корректировать результаты распознавания образов. Если корректно распознан *флажок*, то результатом анализа всего изображения будет означенный экземпляр онтологической структуры понятия *кораблик с флажком*. Его описание может быть синтезировано на ЕЯ и представлять собой различные варианты текстов (в зависимости от априорных требований). Приведем два варианта текста – краткое и подробное:

- *На изображении кораблик с флажком, который расположен в левой верхней четверти экрана.*
- *На изображении представлен кораблик с флажком, который состоит из остова кораблика* (= трапеция с координатами вершин в выбранной системе координат: X1 = 261.09, Y1 = 1084.42; X2 = 1079.87, Y2 = 1067.7; X3 = 957.14, Y3 = 1296.5; X4 = 353.853, Y4 = 1307.62), *мачты кораблика* (= отрезок с координатами: X5 = 728.269, Y5 = 1086; X6 = 728.952, Y6 = 405.722), *паруса* (= параллелограмм с координатами вершин: X7 = 528.52, Y7 = 912.446; X8 = 747.868, Y8 = 908.344; X9 = 840.673, Y9 = 554.967; X10 = 621.324, Y10 = 559.07) *и флажка* (= треугольник с координатами: X11 = 724.831, Y11 = 518.537; X12 = 838.117, Y12 = 480.311; X13 = 724.757, Y13 = 405.722). *Кораблик с флажком расположен в левой верхней четверти экрана.*

3. Система концептуального синтеза изображений

Система концептуального синтеза (плоских) графических изображений является не альтернативой традиционным методам машинной графики, а надстройкой над ними, обеспечивающей дополнительные возможности [Власов и др., 1988].

Задача синтезатора изображений заключается в выполнении на основе онтологических структур графических построений объектов, необходимых для их визуализации, т.е. по полностью или частично означенному концептуальному описанию графической ситуации построить и визуализировать на плоскости отображения графическое изображение.

Концептуальная модель (прикладная онтология) описывает структуры геометрических объектов, составляющих фигуру и связи между их параметрами. Параметры соответствуют общепринятым свойствам фигуры как целого, а также свойствам всех составляющих ее элементов. Например, длина (отрезка), x-координата (точки), размер (угла). Некоторые параметры в описаниях структур могут быть фиксированы, если для всех экземпляров фигуры это значение постоянно. Например, в описании прямоугольного треугольника один из углов равен 90^o. В противном случае значение не определено. Например, в описании равностороннего треугольника конкретное значение длин сторон не фиксировано.

Модель геометрического объекта, у которого всем необходимым вершинам присвоены значения (означены), является описанием конкретного экземпляра фигуры и может быть визуализировано на плоскости отображения.

Задача построения фигуры по имени, структуре и означенным параметрам решается путем вычисления (с учетом закономерностей, присущих данной фигуре и описанных в ее модели) координат вершин (или координат концов аппроксимирующих отрезков) и составления общепринятого в машинной графике описания («в отрезках»).

Программа-планировщик, управляющая расчетом, универсальна по отношению ко всем используемым моделям, а специфичность ее работы для каждой модели геометрического объекта определяется содержанием моделей.

Система должна реагировать на недостаток и противоречивость данных означивания (например, по описанию, заданному пользователем), т.е. невозможность однозначного вычисления экземпляра фигуры. В этом случае она выдает сообщение и список параметров, которые системе не удается вычислить или они противоречивы.

4. Система лингвистического анализа ЕЯ-текстов

Для многих прикладных областей обработки текстов на естественном языке (ЕЯ) лингвистический анализ можно рассматривать как задачу, решающую две основные проблемы: снятие всевозможных неопределенностей в тексте и представление текстовой (явной и неявной) информации на языке прикладной онтологии (модели предметной области).

Задачу лингвистического анализа будем трактовать как преобразование предложения $t_i \in \boldsymbol{T}$ (где $\boldsymbol{T}$ - множество всех предложений ЕЯ) в некоторое описание $m_i \in \boldsymbol{M}$ (где $\boldsymbol{M}$ - множество семантических описаний всех ситуаций в онтологии предметной области), или как отображение Ψ: $\boldsymbol{T} \rightarrow \boldsymbol{M}$, позволяющее по заданному предложению $t_i \in \boldsymbol{T}$ построить адекватное ему описание $m_i \in \boldsymbol{M}$. Это отображение должно устранить неопределенности поверхностного и глубинных уровней ЕЯ (омонимию, омографию, полисемию, неполноту, некорректность и другие), сводя их к однозначному семантическому представлению. Отображение Ψ при этом можно рассматривать как реализацию трех отображений: грамматический анализ, семантическая интерпретация и семантический анализ. Грамматический анализ включает морфологический анализ и синтаксический разбор. Семантическая интерпретация реализует способы "перевода" фрагментов текста во фрагменты прикладной онтологии в зависимости от синтаксических правил (контекстов). Семантический анализ «оформляет» описание ситуации на языке прикладной онтологии.

Морфологический анализ (в случае процедурного воплощения) осуществляет морфологический разбор словоформы на основу и флексию, поиск основы в словаре и по найденной словарной статье приписывание словоформе соответствующих грамматических признаков. Существует некоторое множество морфологических анализаторов, из которых можно использовать любой из них, например, [Мальковский, 1985].

Разнообразие синтаксических анализаторов «поражает воображение», но в основе большинства из них лежит либо дерево составляющих, либо дерево зависимостей или язык расширенных сетей перехода. Среди разработчиков ведутся постоянные споры о преимуществах каждого из направлений. Есть даже попытки соединения первых двух подходов [Гельбух, 1999]. Представляемый ниже подход ближе всего к аппарату расширенных сетей переходов, реализованный в системе ROBOT и в переводчике Systran [Кулагина, 1979].

Семантическая интерпретация практически никем не выделяется в качестве отдельного этапа. Иногда она частично (в пределах поверхностной семантики) «запрятывается» в синтаксический анализ.

Семантический анализ в той или иной степени присутствует только в практических системах доступа на ЕЯ к базам данных [Попов , 1982], [Нариньяни, 1995].

Рассматриваемый в интегральной системе лингвистический анализатор (см. [Хахалин и др., 2006]) содержит два компонента: один базовый, второй расширенный. Базовый компонент транслятора обеспечивает перевод с ЕЯ на язык онтологии полных простых фраз и предложений. Расширенный компонент предназначен для трансляции элементов связного текста и, в частности, осложненных, сложных и эллиптических предложений.

4.2. Базовый алгоритм лингвистического анализа

ЕЯ-предложение t_i будем рассматривать как граф $t_i(x_j^k)$, где x_j^k - элементы предложения (верхний индекс задает порядок слов в предложении). Под элементами подразумеваются словоформы, знаки препинания, скобки, сокращения и другие "вкрапления" в ЕЯ-текст, которые получены после графематического анализа (здесь он не рассматривается).

Полный лингвистический анализ заключается в последовательных морфологическом и синтаксическом анализах, семантической интерпретации и семантическом анализе. Каждый этап лингвистического анализа поддерживается соответствующими моделями. Для морфологического анализа служит морфологическая база данных (морфологический словарь). Синтаксический анализ поддерживается специальной онтологией предметной области "синтаксис ЕЯ", которая строится на языке гиперграфов по тем же принципам, что и прикладная онтология «Планиметрия», только в качестве понятий и отношений выступают элементы синтаксиса естественного языка. Семантическая интерпретация основывается на модели продукционного типа, которая осуществляет "перевод" грамматически разобранных

фрагментов текста во фрагменты прикладной онтологии. А семантический анализ обеспечивается общей с другими системами прикладной онтологией.

Прежде чем рассмотреть алгоритм лингвистического анализа, кратко охарактеризуем каждую из этих моделей.

4.2.1. Грамматическая онтология. Модель грамматики строится по аналогии с прикладной онтологией и образуется совокупностью фрагментов двух типов, которые можно представить в виде гиперграфов. Фрагменты 1-го типа - графы с раскрашенными вершинами и ребрами G_{1i} (X_g, R_{g1}), где X_g — множество элементов онтологии грамматики, включающей слова, грамматические признаки, категории и т. п., а R_{g1} — множество морфологических, родовидовых и структурных отношений («имеет род, число, время, залог ...», «является видом», «входит в структуру» и т.п.). Каждый такой фрагмент представляет собой словоформу со всеми грамматическими признаками, обобщениями и связями с более крупными единицами. Объединение этих фрагментов $G_1 = \bigcup_{k=1}^{n} G_{1i}$ образует морфологическую часть ГМ.

Синтаксическая информация представляется в виде фрагментов 2-го типа, называемых контекстами или контекстными правилами. Контексты образуют иерархическую структуру, которая задается рекурсивно некоторым множеством графов различных уровней. Контекст 1-го уровня определяется как граф G^1_{2j} (X_g, R_{g2}), где $R_{g2} = R_{g1} \cup R_g$, а R_g — система выбранных синтаксических отношений (согласование, управление, следование и т.д.).

Контекст 2-уровня определяется как гиперграф G^2_{2j} ($X_g \cup \{G^1_2\}$, R_{g2}), где $\{G^1_2\} \neq \varnothing$, т. е. граф G^2_{2j} содержит вершины из X_g и хотя бы одну вершину из $\{G^1_2\}$. Тогда множество графов-контекстов k-го уровня определяется выражением $\{G^k_2(X_g \bigcup_{i=0}^{k-1} \{G^1_2\},\ R_{g2})\}$. Объединение фрагментов 2-го типа $G_2 = \bigcup_{i=1}^{n} G^i_2$ образует модель синтаксиса. А вся ГМ есть объединение $G = G_1 \cup G_2$. Другими словами, грамматическая онтология образует некоторый синтаксический гиперграф, в котором присутствуют слова, их обобщения (например, части речи) и контексты, определяющие правила синтаксической сочетаемости. Т.е. эта модель построена по аналогии с тем, как человек, познающий грамматику ЕЯ, декларативно строит систему правил сочетаемости слов (в отличие от процедурного представления с помощью какой-либо формальной грамматики).

Контекстные правила могут задаваться на любом уровне обобщения своих элементов. На уровне словоформ, основ, лексем, всевозможных классов и т.д. Контексты могут представлять шаблоны для выделения в предложениях дат, чисел, имен файлов, географических названий, фамилий и т.п. В контексты могут быть добавлены семантические признаки, лишь бы была возможность их выявления в тексте, например, могут использоваться т.н. ролевые структуры. Сама наполняемость контекстных правил и их номенклатура зависит от разработчика синтаксической части онтологии грамматики и его принадлежности к той или иной лингвистической школе. Фрагмент синтаксической онтологии в виде гиперграфа представлен на рисунке в [Кузин и др., 1989].

4.2.2. Модель семантической интерпретации. Для каждого словарного элемента естественного языка задается "гнездо" продукций вида [Поспелов, 1]:

(i); Q_1; P_1; $A_1 \Rightarrow B_1$; N_1
Q_2; P_2; $A_2 \Rightarrow B_2$; N_2
……...
Q_n; P_n; $A_n \Rightarrow B_n$; N_n

Здесь **(i)** — имя продукции, с помощью которого данная продукция выделяется из множества продукций. В качестве имени может выступать слово (основа), словосочетание, знаки препинания и т.п., отражающие суть данной продукции.

Элемент **Q** характеризует сферу применения продукции - тематику текста. Тема текста может динамически определяться известными статистическими методами.

Основным элементом продукции является ее ядро **A ⇒ B**. Интерпретация ядра продукции может быть различной. Обычное прочтение ядра - ЕСЛИ **A**, ТО **B**. Более сложные конструкции ядра допускают в правой части альтернативный выбор, например, ЕСЛИ **A**, ТО **B1**, ИНАЧЕ **B2**. В нашем случае **A** - некоторое упорядоченное множество контекстных правил онтологии грамматики. **B** - соответствующее **A** множество понятий, отношений или их сочетаний из модели семантики (прикладной онтологии). Для не интерпретируемых элементов текста (знаки препинания, предлоги, союзы и т.п.) множество **B** = ∅.

Элемент **P** есть условие применимости ядра продукции, и определяется принадлежностью анализируемого слова к определенной части речи.

Элемент **N** описывает постусловия продукции. Они актуализируются в том случае, если ядро продукции реализовалось. Постусловия описывают действия и процедуры, которые выполняются после реализации **B**.

Параметр $n \geq 1$ характеризует множественность интерпретации элемента текста.

Модель интерпретации представляет собой словарь системы, где заданы способы "перевода" элементов текста в понятия в зависимости от синтаксических правил.

4.2.3. Семантическая модель или прикладная онтология. Модель семантики естественного языка приравнивается к прикладной онтологии, которая описана в разделе 1.3. Имена концептов и отношений могут ассоциироваться со словами ЕЯ, но выполняют номинативную функцию и используются для упрощения и удобства процессов разработки, наполнения и отладки модели. На языке прикладной онтологии информация задается явно, даже если она неявно представлена в тексте. Например, если в тексте есть словосочетание *красный шар*, то на языке онтологии эта ситуация будет представлена в виде тройки (*шар*) (*имеет_цвет*) (*красный*).

4.2.4. Грамматический анализ. Морфологический анализ на основе декларативно или процедурно заданного морфологического словаря приписывает каждому слову в предложении в общем случае множественные грамматические характеристики. Для рассмотрения синтаксического разбора введем некоторые определения.

Определение 1. Фрагментом f_{ij} графа предложения t_i будем называть его подграф $t_{ij} \subseteq t_i$ k-изоморфный определенному графу контекста G_{2i}^{k}.

Например, в предложении *мальчик спит непробудным сном* фрагментом может быть словосочетание (*мальчик спит*), которое выделено с помощью контекста согласования глагола с существительным.

Определение 2. Два фрагмента f_{ik} и f_{im} одного предложения t_i будем называть связными или пересекающимися, если пересечение графов $t_{ik} \cap t_{im} \neq \varnothing$.

Например, в предложении *стены древнего города* можно выделить два пересекающихся фрагмента (*стены города*) и (*древнего города*).

Определение 3. Фрагмент f_{ij} будем называть вложенным во фрагмент f_{ik}, если граф t_{ij} является подграфом графа t_{ik}.

Определение 4. Фрагмент f_{ij} будем называть изолированным, если $t_{ij} \bigcap \bigcup_{k=1}^{n} t_{ik} = \varnothing$ и $j \neq k$.

Определение 5. Предложение t_i будем считать полностью покрытым фрагментами, если каждое слово $x_j^k \in t_i$ входит, по крайней мере, в один из фрагментов из $\{f_{ij}\}$, где $\{f_{ij}\}$ - множество (полного) покрытия.

Примером полного покрытия может быть фрагментация предложения (*Хозяйка мыла*) (*грязное стекло*), где первый фрагмент выделен с помощью контекстного правила ГК4, а второй - правила согласования (прил.) и (сущ.).

Определение 6. Будем говорить, что множество фрагментов $\{f_{in}\}$ образует связную структуру фрагментов для t_i, если для любого связного подмножества этого множества справедливо

$$\bigcup_{j=1}^{k} t_{ij} \bigcap \bigcup_{l=k+1}^{n} t_{il} \neq \varnothing, \text{ где } k = \overline{1, n-1}.$$

Определение 7. Предложение t_i будем считать полностью разобранным, если оно полностью покрыто фрагментами и множество покрытия $\{f_{in}\}$ образует связную структуру фрагментов для t_i.

Примером полного и связного покрытия может быть разбор (*Хозяйка мыла стекло*) (*грязное стекло*).

Синтаксический разбор предложения t_i будет состоять в поиске такого множества фрагментов $\{f_{ij}\}$, которое полностью покрывает t_i и образует при этом связную структуру фрагментов. В остальных случаях можно говорить о частичном синтаксическом разборе.

Процесс полного разбора носит итеративный характер. Он включает в себя: выбор контекстов-гипотез из множества $\{G_{2i}^{k}\}$, сопоставление этих контекстов с t_i и выделение связных фрагментов.

Процесс поиска и выбора релевантных t_i контекстов определяется $\{x_j^k\}$ и удачно сопоставленными контекстами. При этом привлекается информация из $G_{11}(X_g, r_1)$ и $G_{12}(X_g, r_2)$, где r_1 - отношение "является видом", а r_2 - отношение "входит в структуру". При удачно сопоставленном контексте G_{2j}^{i} уровня i очередная гипотеза выбирается из множества контекстов уровня $(i+1)$ на основе данных о вложенности.

В результате успешного сопоставления выделяется фрагмент f_{ij}. Если для него справедливо $t_{ij} \cap \bigcup_{k=1}^{l} t_{ik} \neq \varnothing$, где l - число связанных фрагментов для t_i, то он дополняет связную структуру фрагментов $\{f_{il+1}\}$. После чего разбор продолжается для слова $x_i^{k+m} \notin \{f_{il+1}\}$, где m - количество слов, принадлежащих f_{ij}. Если f_{ij} - изолированный фрагмент, то разбор проводится для слова $x_i^{k+1} \notin \{f_{il+1}\}$, т.е. для следующего слова, не входящего в связную структуру фрагментов. А сам изолированный фрагмент запоминается с целью его возможного включения в структуру фрагментов в последующих итерациях.

Таким образом, процесс повторяется до получения полного разбора. Результатом этого процесса будет граф разбора τ_i, образованный графами связных фрагментов и данными о сопоставлении контекстов с t_i.

Пример синтаксического разбора…

Если в процессе разбора невозможно получить полное и связное покрытие, то относительно входного предложения можно предположить, что разбор происходил по ложной ветке (в предложении есть неопределенности), что предложение является неполным или оно некорректно с точки зрения заданного синтаксиса. Например, если в предложении *хозяйка мыла грязное окно* выделены фрагменты (*хозяйка мыла*) (слово *мыла* рассматривается как существительное) и (*грязное окно*), но они не удовлетворяют критерию связности, то грамматический анализатор пытается рассмотреть омоним слова *мыла* в качестве глагола. В этом случае получаем связные фрагменты (*хозяйка мыла окно*) и (*грязное окно*).

4.2.5. Семантическая интерпретация. Языки синтаксиса и семантики различны, а в общем случае различна и фрагментация в них. Лексику М-языка составляет множество проблемных понятий и отношений, которые обозначим через p_i. Введем еще пару определений для понятия *подстрочника* предложения t_i.

Определение 8. Некоторое частично упорядоченное множество $m_i^o = (\{p_i^1\} \cup \{p_{ik}^1\}, \{p_i^2\} \cup \{p_{il}^2\}, ..., \{p_i^n\} \cup \{p_{im}^n\})$ образует в лексике языка прикладной онтологии подстрочник $t_i \in T$, если множества $\{p_i^l\}$ - множество значений в М-языке составляющих t_i (слов и словосочетаний); а $\{p_{ij}^l\}$ - прессупозиции, определяемые этими составляющими; n-число проинтерпретированных составляющих.

Определение 9. Подстрочник является однозначным, если для любого $l = \overline{1,k}$ справедливо $|\{p_i^l\}| = 1$, т.е. он имеет вид $m_i^o = (p_i^1, p_i^2, ..., p_i^n)$, где $k \leq n$ (здесь прессупозиции опущены). Число $(n\text{-}k)$ характеризует количество неопределенностей (омонимий и др.), снятых относительно множеств $\{p_i^l\}$.

Тогда семантическая интерпретация будет состоять в получении по заданному графу грамматического разбора τ_i однозначного подстрочника m_i^o. Процесс интерпретации включает в себя построение с помощью системы продукций подстрочника по частям, соответствующим фрагментам f_{ij} и снятие неопределенностей в частях подстрочника.

Для элементов $x_j^k \in f_{ij}$ каждого фрагмента f_{ij} по модели интерпретации получаем часть m_i^o. Реализуя итеративно процесс для всех f_{ij}, получаем весь подстрочник. Снятие неопределенностей в подстрочнике осуществляется операцией пересечения множеств $\{p_i^k\}$. Если для определенных $m \neq k$ при последовательном пересечении справедливо $|\{p_i^k\} \cap \{p_i^i\}| = 1$, где $\{p_i^k\}$ и $\{p_i^i\}$ принадлежат части подстрочника, соответствующей одному фрагменту f_{ij}, то неопределенность снимается и одно множество исключается из m_i^o, второе множество заменяется единственным элемент пересечения, а прессупозиции для этих двух множеств объединяются. Если невозможно получить однозначный подстрочник, то предложение считается семантически некорректным с точки зрения интерпретации.

Однозначный подстрочник рассматривается аналогично графу предложения t_i, поэтому все, что справедливо для t_i, справедливо и для m_i^o.

4.2.6. Семантический анализ предложений текста. В процессе интерпретации в семантической модели (= прикладной онтологии) получены фрагменты семантического описания ситуации, представленной в предложении на ЕЯ. Теперь остается «связать» эти фрагменты в некое целостное описание на языке прикладной онтологии, одновременно проверив полноту и корректность такого представления. Эту задачу выполняет семантический анализ, который выполняется по полной схеме (аналогичной полному грамматическому анализу), в результате которого получается структура на языке прикладной онтологии, соответствующая ситуации.

Описанный выше алгоритм лингвистического анализа предназначен для простых полных предложений. Трансляция элементов связанного текста (эллиптических, осложненных и сложных предложений) осуществляется расширенным компонентом анализатора.

4.3. Расширенный алгоритм лингвистического анализа

Анализатор любое сложное предложение разбивает на простые фразы, и каждая фраза транслируется базовым компонентом. Разбор сложных и эллиптических предложений необходимо рассматривать вместе, поскольку «разбивка» осложненных и сложных предложений часто порождает неполные фразы, которые необходимо связывать с другими фразами предложения. Кроме этого существует самостоятельная задача разбора отдельных неполных ЕЯ-предложений.

4.3.1. Трансляция эллиптических предложений. Эллипсисы характеризуются неполнотой. Формально можно предположить, что эллиптичность проявляется как на уровне синтаксиса, так и на уровне семантики.

В рассматриваемом методе анализа синтаксическим эллипсисом будет такое правильно построенное предложение t_i, для которого справедливо

$$\exists x_i^k \left(x_i^k \in t_i \,\&\, x_i^k \notin \{f_{ij}\}\right),$$

т.е. в предложении t_i существует, по крайней мере, одно такое слово x_i^k, для которого нельзя найти фрагмент f_{ij}, расширяющий связную структуру синтаксических фрагментов $\{f_{ij}\}$ (предложение считается синтаксически разобранным полностью, если оно «покрыто» фрагментами и эти фрагменты образуют связную структуру; определения полного и связного покрытия выше). Для семантических эллипсисов существует аналогичное условие.

Обработка эллипсисов включает два этапа: восстановление их до полных фраз за счет дискурса и трансляция восстановленных фраз с помощью базового компонента. В качестве дискурса используется локальный дискурс (для сложного предложения) или глобальный дискурс (для простых неполных предложений). Отсутствие дискурса или невозможность корректного восстановления эллипсиса характеризует нарушение связности ЕЯ-текста или неполноту соответствующей модели.

Восстановление эллипсисов включает поиск аналогичных фрагментов дискурса и эллипсиса и добавление из дискурса в эллипсис недостающих элементов с их возможной коррекцией. Если рассмотреть задачу доступа к базам данных на ЕЯ (например, в кадровой задаче) и в качестве текста для ЛТ задать последовательность предложений *Сколько сотрудников отдела маркетинга получают зарплату больше 100 долларов?* и *Отдела снабжения?*, то второе предложение будет синтаксическим эллипсисом. Сопоставление дискурса и эллипсиса даст соответствия *отдела* ↔ *Отдела* и *маркетинга* ↔ *снабжения*, а добавления из дискурса в неполную фразу дадут в результате полностью восстановленный эллипсис в виде:

Сколько сотрудников отдела снабжения получают зарплату больше 100 долларов?

4.3.2. Трансляция сложных предложений. Трансляция сложных предложений основана на базовом компоненте для полных фраз и на схеме трансляции эллипсисов для неполных фраз. Трансляция включает следующие этапы: разбивка сложного предложения на фразы по структурным признакам «усложнения» (см. ниже); получение текущей фразовой структуры предложения с последующим ее уточнением; итеративная трансляция каждой выделенной фразы и «сочленение» описаний на языке прикладной онтологии в общую структуру на основе окончательной фразовой структуры ЕЯ-предложения.

Для правильно построенных осложненных и сложных предложений всегда существуют признаки «усложнения»: союзы, союзные слова, знаки препинания и т.п. Для каждого ЕЯ существует ограниченное множество типов сложных предложений. Все это позволяет внести в грамматическую модель транслятора понятия и структуры «связок», необходимые для разбивки осложненных и сложных предложений. Для русского языка (например, по [Розенталь и др, 1995]) подобное множество состоит из примерно 250-300 структур. Все структуры естественно «погружаются» в некоторую связную модель, в которой существуют отношения типа «является видом», «входит в структуру» и т.д., с помощью которых можно осуществлять поиск и сопоставление структур «связок» с входным предложением. Элементами структур связок могут быть конкретные словоформы, лексемы, части речи, пунктуационные знаки и различные их сочетания, между которыми существуют синтаксические и геометрические отношения. Примеры изолированных структур связок приведены на рис. 6.

Каждая структура имеет свое уникальное в данной модели имя и может обладать некоторыми свойствами (свойство, характеризующее вид усложнения, например, сложноподчиненное определительного типа или вводная конструкция и т.п.).

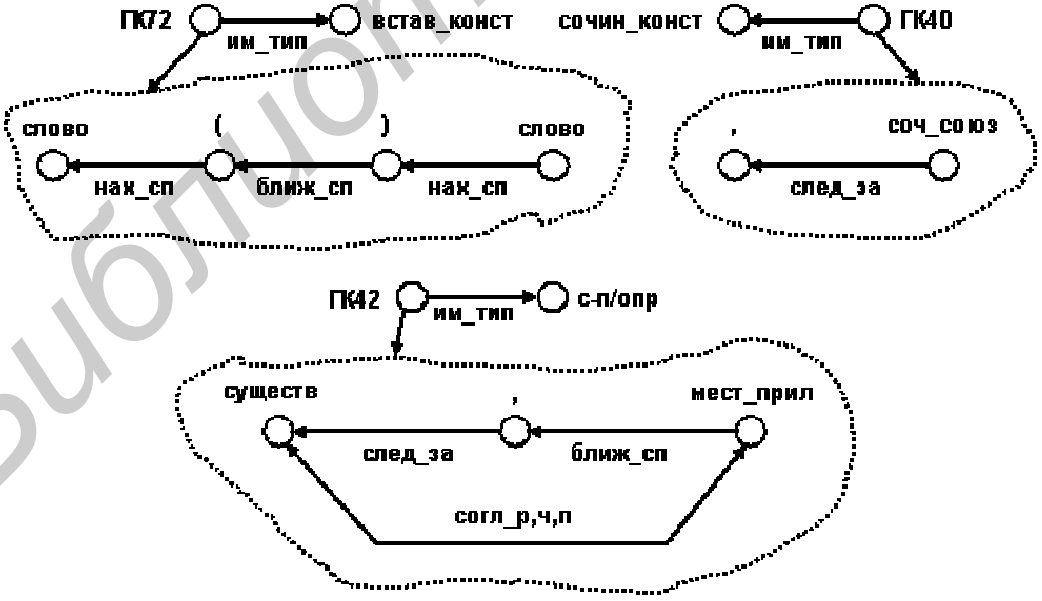

Рисунок 6 - Примеры структур «связок» фраз для декомпозиции сложных предложений

На рисунке приняты следующие обозначения: понятия «существ» – существительное, «соч_союз» – любой сочинительный союз, «мест_прил» – любой элемент из класса местоименных прилагательных (остальные понятия очевидны); отношения «след_за» – следует

за, «ближ_сп» – ближайший справа, «нах_справа» – находится справа и «согл_р,ч,п» – одновременное согласование элементов структуры и соответствующих слов в предложении в роде, числе и падеже.

Процесс разбивки включает следующие этапы: выбор структур «связок» для разбивки; сопоставление этих структур «связок» с ЕЯ-предложением; выделение связующих частей предложения и итеративное выделение фраз. Положительный результат сопоставления структуры связки с предложением дает возможность выделить признак типа рассматриваемого сложного предложения.

Такая разбивка сложного предложения на фразы учитывает множественность взаимосвязей отдельных слов с разными фразами, т.е. некоторое слово может попадать сразу в несколько фраз. Например, если задано предложение *Найти площадь равностороннего треугольника, катет которого равен 15 см, а высота – 12 см.*, то в результате разбивки получим три фразы: *Найти площадь равностороннего треугольника | треугольника, катет которого равен 12 см | высота – 15 см*, где слово *треугольника* будет присутствовать как в первой фразе, так и во второй. Кстати, вторая фраза при трансляции будет эквивалентна фразе *катет треугольника равен 12 см* (на основе анафорических преобразований). Исходя из структур связок, используемых для разбивки, получаем фразовую структуру всего предложения: сложноподчиненное предложение определительного типа, где подчиненная часть является сложным сочинением. Третья фраза эллиптична и может быть восстановлена до фразы *высота треугольника равна 15 см* за счет второй на основе фразовой структуры предложения. Отметим, что полученная фразовая структура носит характер не дерева (как в системе составляющих или в дереве зависимостей), а более сложного, но взаимосвязанного представления.

Каждая фраза (после восстановления эллипсиса в третьей фразе) будет оттранслирована базовым компонентом транслятора с учетом связей между элементами фраз. В результате будет получено описание ситуации, представленной ЕЯ-предложением, на языке модели проблемной среды в виде, как это показано на рисунке 7.

Рисунок 7 - Представление ситуации, описанной сложным предложением, на языке онтологии

4.3.3. Обработка семантически неполных предложений. Семантические эллипсисы восстанавливаются в модели проблемной среды по схеме, описанной выше. Например, если на ЕЯ задана пара предложений типа: *Задан прямоугольный треугольник с высотой 5 см и катетом 10 см. Найти площадь треугольника*, то второе предложение является полным с точки зрения синтаксиса, но оно семантически неполно, поскольку не ясно, к какому понятию в онтологии относится структура, представленная на рисунке 8.

Рисунок 8 - Пример представления семантического эллипсиса

Если эту структуру отнести непосредственно к понятию «треугольник», то общий результат для этого текста не даст связное описание. Обработка семантического эллипсиса заключается в «привязке» к тому описанию в онтологии, которое было получено при трансляции первого

предложения, т.е. к понятию «прямоугольный треугольник». В результате восстановления будет получено описание, которое аналогично представленному на рисунке 7.

4.4. Настраиваемая последовательность обработки при анализе

Для многих задач, где используется лингвистический анализ ЕЯ-текста, нет необходимости проводить полный и последовательный синтаксический анализ предложений. Например, для задач доступа к БД, для поиска по ключевым словам и т.п. Напротив, для других задач необходим полный синтаксический анализ. Например, для машинного перевода, для семантического представления документов и т.д. То есть для множества прикладных задач существует спектр полноты и последовательности лингвистического анализа. Это зависит от решаемой задачи, проблемной области, онтологии, тематической однородности текста и т.д. Поэтому хотелось бы иметь такую архитектуру анализатора, которая подключала бы соответствующие компоненты анализа в зависимости от сложности самого текста. Предлагаемая структура лингвистического анализатора позволяет реализовать подобную схему. Инициализация процесса анализа в такой схеме исходит от интерпретации.

Схема с настраиваемой последовательностью обработки может быть представлена следующим образом. Пусть на вход анализатора поступает ЕЯ-текст. Первое выделенное из текста предложение проходит морфологический анализ. Далее выбирается первое слово с его морфологическими признаками и запускается процесс интерпретации этого слова. Если слово имеет единственную интерпретацию, то осуществляется переход к другому слову. В противном случае выбирается первое контекстное правило из упорядоченного множества правил, заданного в гнезде продукций данного слова. Выбранное правило сопоставляется с морфологической структурой предложения.

На основе результатов сопоставления осуществляется интерпретация фрагмента предложения либо переход к другому контекстному правилу. И так до тех пор, пока не будет выделен фрагмент предложения либо не будет исчерпан список контекстных правил. В последнем случае предложение считается некорректным.

При успешном сопоставлении контекстного правила с предложением получаем частичный синтаксический разбор предложения, т.е. фрагмент разбора с некоторыми уже проанализированными словами дополнительно к выделенному слову.

Далее продолжается интерпретация слов, которые не вошли в какой-либо фрагмент разбора. Этот процесс итеративно продолжается до тех пор, пока не будут проинтерпретированы все слова предложения. Требования полного покрытия и связности фрагментов, необходимые при полном синтаксическом разборе, не являются обязательными при частичном разборе. Если при этом полученное полное покрытие предложения фрагментами не удовлетворяет критерию связности, то его несвязность может быть обнаружена на последующем семантическом этапе анализа, который проводится по полной схеме.

Управление степенью полноты синтаксического разбора осуществляется семантическим интерпретатором в зависимости от степени неопределенности в самом предложении.

5. Система лингвистического синтеза ЕЯ-текстов

Задача синтеза текстов на ЕЯ заключается в генерировании текста по структурным описаниям онтологических представлений. Полный синтез фраз ЕЯ предполагает этапы семантического синтеза, синтаксической интерпретации, синтаксического синтеза, морфологического синтеза и форматирования (графематический синтез).

Следует отметить, что синтез ЕЯ-предложений с уровня онтологии еще пока недостаточно проработанная тематика. Обычно при синтезе используют «шаблонный» метод, когда в систему закладываются заранее определенные шаблоны (ответы, вопросы, пояснения) с заранее определенными «лакунами». И по мере появления необходимости синтезировать ЕЯ-предложение система находит соответствующий шаблон и вставляет на место «лакун» необходимые параметры, которые могут представлять собой числа и слова.

Примером, где используется полная схема синтеза ЕЯ-предложений с уровня модели предметной области, может служить система ПОЭТ [Попов, 1982].

Синтез с синтаксического (поверхностно-семантического) уровня достаточно хорошо разработан и описан в литературе, например, в [Апресян и др., 1988].

Библиографический список

[Апресян и др., 1988] Апресян Ю. Д. и др. Лингвистическое обеспечение системы ЭТАП-2 - М.: Наука, 1988.

[Байков и др., 1980] Байков А. М., Кузин Е. С., Хахалин Г. К., Шамис А. Л. Управляемый контекстом целенаправленный анализ изображения на основе использования семантической сети // Вопросы радиоэлектроники, сер. ОТ, вып. 1. – М., 1980, С. 50-58.

[Башмаков и др., 2006] Башмаков И. А., Башмаков И. А. Интеллектуальные информационные технологии. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2006.

[Визинг, 2007] Визинг В. Г. О раскраске инциденторов в гиперграфе // Дискретный анализ и исследование операций. – 2007. - Серия 1. - Том 14. - № 3. – С. 40–45.

[Власов и др., 1988] Власов А. В., Аредова И. И. Экспериментальная система синтеза графических изображений по их описанию в терминах геометрических понятий // Материалы конференции "Развитие интеллектуальных возможностей современных и перспективных ЭВМ". - М.: МДНТП, 1988, С. 123-132.

[Гаврилова, 2006] Гаврилова Т. А. Формирование прикладных онтологий // Труды XX национальной конференции по Искусственному Интеллекту с международным участием – КИИ-2006, т. 2 – М.: Физматлит, 2006.

[Гельбух, 1999] Гельбух А. Между смыслом и текстом // Труды Труды Международного семинара "Диалог'2009" по компьютерной лингвистике и ее приложениям. – М., 1999, С. 47–55.

[Зыков, 1974] Зыков А. А. Гиперграфы // Успехи математических наук. – 1974. - Т. 29. - вып. 6. - С. 89–154.

[Кузин и др., 1989] Кузин Е.С., Ройтман А.И., Фоминых И.Б., Хахалин Г.К. Интеллектуализация ЭВМ. Книга 2 серии "Перспективы развития вычислительной техники". - М.: Высшая школа, 1989.

[Кулагина, 1979] Кулагина О. С. Исследования по машинному переводу, М.: Наука, 1979.

[Курбатов и др., 2010] Курбатов С.С., Найденова К.А., Хахалин Г.К. О схеме взаимодействия в комплексе «анализ и синтез естественного языка и изображений» // Труды XII национальной конференции по Искусственному Интеллекту с международным участием – КИИ-2010 (Тверь, 20-24 сентября 2010) в 4-х томах, т. 1. – М.: Физматлит, 2010, С. 234-242

[Курбатов, 2010] Курбатов С. С. Высокоуровневые эвристики для автоматизированного Формирования базы знаний // Труды XII национальной конференции по Искусственному Интеллекту с международным участием – КИИ-2010 (Тверь, 20-24 сентября 2010) в 4-х томах, т. 1. – М.: Физматлит, 2010, С. 231-239.

[Мальковский, 1985] Мальковский М. Г. Диалог с системой искусственного интеллекта. - М.: МГУ, 1985.

[Найденова и др., 2008] Найденова К. А., Невзорова О. А. Машинное обучение в задачах обработки естественного языка: обзор современного состояния исследований // Известия Казанского Университета, №1, 2008 г., С. 3-24.

[Нариньяни, 1995] Нариньяни А. С. Проблема понимания запросов к базам данных решена // Труды Международного семинара по компьютерной лингвистике. Казань. 1995, С. 206-215.

[Попов , 1982] Попов Э. В. Общение с ЭВМ на естественном языке.- М.:Наука, 1982.

[Розенталь и др., 1995] Розенталь Д. Э., Голуб И. Б., Теленкова М. А. Современный русский язык. — М.: Международные отношения, 1995.

[Хахалин и др., 2006] Хахалин Г. К., Воскресенский А. Л. Контекстное фрагментирование в лингвистическом анализе // Труды X национальной конференции по Искусственному Интеллекту с международным участием – КИИ-2006. М.: Физматлит, 2006, с. 479-488.

[Хахалин, 2009] Хахалин Г.К. Прикладная онтология на языке гиперграфов // Труды второй Всероссийской Конференции с международным участием "Знания-Онтологии-Теории" (ЗОНТ-09). Новосибирск. – 2009. - С. 223-231.

[Denisov et al., 1991] Denisov, D. A., Plaksin M. V. Object Synthesis in Remote Sensing Imagery Understanding // International Journal of Systems and Technology. - 1991. - v. 3. - P. 249-256.

[MyScript Notes, 2007] Режим доступа: http:// www.visionobjects.com/en/webstore/myscript-studio/description/. – [Электронный ресурс].

[Naidenova, 2004] Naidenova, X.A. Model of Common Sense Reasoning Based on the Lattice Theory. Abstracts of the Conference “Mathematical Methods for Learning - 2004. Advances in Data Mining and Knowledge Discovery”, June 21-24 2004, Como, Italy. P. 36- 39.

[Tandareanu et al., 2003] Nicolae Tandareanu and Mihaela Ghindeanu. Image Synthesis from Natural Language Description. http://www.inf.ucv.ro/~ntand/Publications/ntand-2003+MG2.pdf

[Wang et al., 2009] Wang, J., Markert, K., Everingham M. Learning Models for Object Recognition from Natural Language Descriptions. 20 British Machine Vision Conferences (BMVC2009), Sept 2009.