Автоматизация процесса извлечения онтологической информации из вербальных терминологических словарей (на примере терминологического словаря задачи межзвездного поглощения)

© К.К. Боярский[1], Е.А. Каневский[1], Г.В. Лезин[1], Л.А. Калиниченко[2], Н.А. Скворцов[2]

[1]Санкт-Петербургский экономико-математический институт РАН

[2]Институт проблем информатики РАН, г. Москва

kirill@eu.spb.ru, {kanev, lezin}@emi.nw.ru, {leonidk, nskv}@ipi.ac.ru

Аннотация

Рассматривается задача построения онтологической модели предметной области по ее спецификации, заданной терминологическим словарем. Алгоритмы извлечения онтологической информации из терминологического словаря задаются набором продукционных правил, применяемых к результату семантико-синтаксического анализа дефиниций словаря. Разработана программа интерпретации таких правил и проведен эксперимент по разработке правил и их применению для небольшого узкоспециального словаря. В докладе приводится предварительный анализ результатов эксперимента.

1 Введение

Задача автоматического извлечения и формализации знаний, содержащихся в терминологических словарях, привлекает внимание исследователей уже достаточно давно [1 – 3]. Терминологические словари, и общие энциклопедические, и ориентированные на описание отдельных предметных областей, и узкоспециализированные для той или иной области человеческой деятельности, в совокупности образуют обширнейший свод знаний людей о мире. Информация в словарях так или иначе структурирована, тексты толкований терминов, прошедшие редакторскую экспертизу, как правило, достаточно строго соответствуют нормам естественного языка. Референциальная связность текстов разных определений проявляется, главным образом, на уровне использования общей терминологии. Тексты словарей, в основном, доступны для применения современных методов прикладного лингвистического анализа.

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

Терминологические словари по своей природе онтологичны. Автоматическое выявление структурных взаимосвязей между терминами, как явно заданных в словаре их определениями, так и скрытых, выявляемых в результате анализа явно заданных связей, может быть одинаково полезным как при создании нового словаря [4], так и при его использовании при формировании и пополнении онтологий [3].

В рамках исследовательского проекта создания новой информационной технологии решения задач над множеством интегрируемых неоднородных распределенных информационных ресурсов [5] предлагается новое направление использования терминологических словарей – как источников информации для построения исходных спецификаций предметных областей. Конечной целью построения спецификации является получение концептуальной модели предметной области. В этой модели наряду со статичными связями понятий, свойственными онтологической модели, фиксируются также и методы работы с понятиями, описывающие поведение экземпляров понятий в различных условиях. Речь идет о переходе от онтологической модели предметной области к модели, описывающей предметную область в терминах абстрактных типов данных [6]. Явно выявляется последовательность действий:

- определение терминологического словаря, специфицирующего термины предметной области;
- построение онтологической модели предметной области по результатам анализа терминологического словаря и извлечения из него онтологической информации;
- преобразование полученной онтологической модели в концептуальную модель.

Для проверки жизнеспособности предложенного подхода был проведен сравнительно небольшой эксперимент. В качестве экспериментальной была выбрана достаточно узкая астрономическая задача мезвездного поглощения. Задача специфицировалась набором из 68 астрономических терминов. Далее вручную были построены онтологическая и выводимая из нее концептуальная модели задачи. Сле-

дующим шагом эксперимента была разработка программы, автоматизирующей процесс извлечения онтологической модели из текстов терминологического словаря. Доклад посвящен описанию методов и оценке результатов этого шага в общем эксперименте.

На данном этапе исследований мы наложили довольно существенные ограничения на анализируемые тексты:

а) сочли возможным ограничить определение одним предложением и исключили из анализируемых предложений анафорические отношения;

б) не анализируем референциальный статус термина в текстах, считая «по определению», что термином всегда обозначен класс, причем класс, никак не связанный с общей системой понятий естественного языка; в этих условиях мы считаем, что полный список классов целевой онтологии образован фактически заданным перечнем определяемых терминов и наша задача – выявить и конкретизировать отношения между терминами;

в) общий контекст терминологического словаря образован исключительно списком определяемых терминов; каждое из вербальных определений анализируется независимо от содержания других определений; получаемый формальный результат анализа не зависит от порядка обработки определений;

г) в качестве формального языка для представления целевой онтологии мы используем OWL в упрощенном L-диалекте [7].

Далее будут рассмотрены: общий подход к решению задачи; семантика правил, используемых в процессе анализа определений; основные результаты.

2 Общий подход

Четко выделяются три этапа решения задачи.

А. Семантико-синтаксический анализ исходного варианта текстов вербальных определений. В результате каждому из предложений текста сопоставляется дерево семантико-синтаксических связей между словами предложения (дерево разбора). На этом же этапе выделяются словосочетания, образующие термины терминологического словаря. Выделенные словосочетания сливаются в узлы дерева. В результате каждому из вхождений термина в разбираемое предложение, независимо от количества образующих этот термин слов, в дереве разбора соответствует один узел.

Б. Разработка системы правил, вычленяющих из текста вербального определения онтологически значимую информацию. В принципе общий комплекс правил может содержать как правила достаточно общего плана, использование которых практически не зависит от особенностей конкретной предметной области, так и правила, ориентированные на контекст конкретной предметной области анализируемого терминологического словаря. И если первые могут быть разработаны однократно (хотя и они могут потребовать подстройки в конкретных случаях использования), то вторые приходится разрабатывать и уточнять для каждой новой предметной области. Доступность системы правил для правки (в нашем случае – разработчиком онтологии) – одна из фундаментальных особенностей всех программ, ориентированных на выделение значимой информации из текстов на естественном языке.

В. Применение правил к текстам вербальных определений терминологического словаря. Конечным результатом этой работы является формальный текст онтологии, связывающей исходный набор терминов в сеть отношений между терминами. Полученный текст представлен на стандартом формальном языке (на языке OWL) и может быть введен в стандартный онторедактор для последующего анализа и правки.

Получение конечного результата связано выполнением большого объема итераций:

– правится исходный текст определений для получения приемлемой интерпретации его имеющимися правилами;

– уточняются и пополняются правила интерпретации текстов, т. е. фактически ведется отладка правил.

Общая структура программной системы, реализующей рассмотренный подход, представлена на рис. 1. В основе системы лежит семантический словарь русского языка, поставляющий полную (морфологическую, синтаксическую и семантическую) информацию о лексических значениях слов. С семантическим словарем сопряжен семантико-синтаксический анализатор русских предложений. Мы в нашей работе используем словарь и анализатор, разработанные В.А. Тузовым [8]. Нужно отметить, что классификатор лексических значений слов, разработанный В.А. Тузовым, не вполне соответствует современным онтологическим требованиям, поскольку разрабатывался прежде всего для нужд семантико-синтаксического анализа предложений. В связи с этим в нашей работе лексический классификатор В.А. Тузова дополняется лексически согласованным с ним фрагментом универсальной онтологии естественного языка, структурно аналогичной онтологии, разрабатываемой в проекте DOLCE [9].

Результат работы семантико-синтаксического анализатора, а также текстовое представление результата как исходный материал для последующей обработки лучше всего рассмотреть на примере.

Пример 1. На рис. 2 показано визуальное представление дерева разбора, полученное в результате семантико-синтаксического анализа предложения

«**Галактическая широта** – одна из двух **галактических координат**, измеренная от плоскости **галактики** к **объекту**».

Приведенное на рис. 2 дерево разбора отображает также и результаты постсинтаксической обработки: выделены узлы, которым соответствуют термины исходного терминологического словаря; предлоги слиты с опорными словами предложных групп.

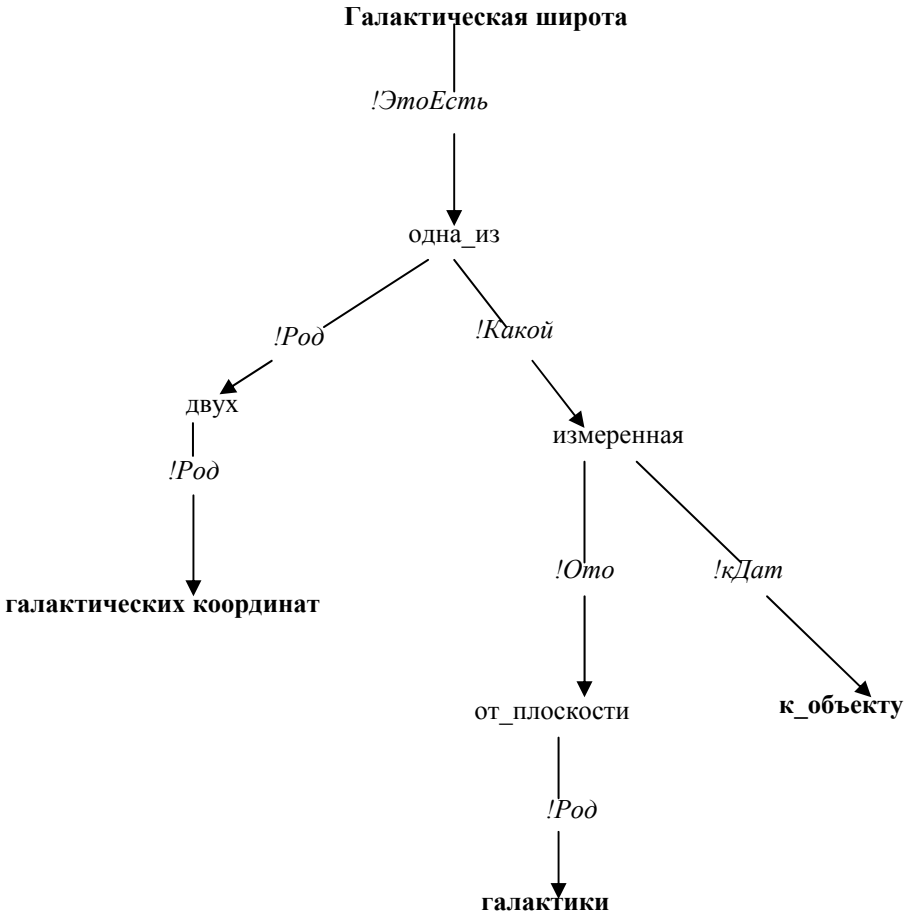

Рис. 2. Пример визуального представления дерева разбора

Набор триплетов, представляющий результат разбора (см. рис. 2):

«**Галактический широта»** *ЭтоЕсть* «одна_из».
«одна_из» *Какой* «измеренная»; *Род* «двух».
«двух» *Род* «**Галактический координата**».
«измеренная» *Ото* «от_плоскости»; *кДат* «**к_объекту**».
«от_плоскости» *Род* «**галактика**».

Кавычками выделены константные значения имен узлов дерева разбора, курсивом – имена синтаксических связей, жирным шрифтом – термины.

В данном материале мы используем простейшую, интуитивно понятную форму записи триплетов, отказавшись от ряда возможностей, предоставляемых стандартными языками описания сетей (например, RDF [10]).

Примечание к примеру:

– словосочетание «одна_из» в виде единого узла сформировано анализатором В.А. Тузова;

– словосочетания «от_плоскости» и «к_объекту» получены в результате постсинтаксической обработки результата синтаксического анализа.

Ключевым моментом наших исследований стал выбор метода извлечения онтологической информации из дерева разбора. Решающим ограничивающим фактором здесь является максимальная «открытость» алгоритма извлечения, возможности его пополнения и уточнения разработчиками конкретных онтологий на этапах практического создания терминологического словаря.

В результате мы остановились на методе, сутью которого является многошаговая трансформация исходного дерева разбора в дерево, представляющее собой фрагмент онтологии терминологического словаря. Трансформация подчинена системе правил. Каждый шаг трансформации является результатом применения одного правила из общего списка.

3 Семантика правил

Трансформационное правило имеет вид продукции и состоит из двух частей:

– решающей части, которая содержит условия применимости правила;

– исполнительной части, содержащей последовательность действий по преобразованию текущего вида и (или) состояния дерева разбора в новый вид (состояние).

Решающая часть правила содержит:

– указание опорного узла дерева разбора, в контексте которого возможно применение данного правила;

– указание, если это необходимо, контекста этого опорного узла на дереве разбора, для которого данное правило может считаться подходящим;

– указание, если это необходимо, свойств (морфологических и семантических), которым должны удовлетворять узлы (все или некоторые) из заданного контекста.

Деревом разбора представлены подчинительные связи между словами предложения. В этих условиях действует общий принцип, согласно которому семантика слова, сопоставленного узлу дерева разбора, должна устанавливаться только после того, как установлена семантика подчиненных ему слов. Соответственно, направление анализа дерева разбора – от листьев к корню.

Введены два принципиально разных вида правил:

– T-правила, применяемые к очередному не обработанному исходящему триплету опорного узла дерева разбора;

– N-правила, применяемые к опорному узлу только после того, как все исходящие триплеты этого узла оказались обработанными T-правилами.

Синтаксически правило оформляется в виде именованного блока информации, атрибутом которого определяется T- или N-тип этого правила:

<*имя_правила* ТИП={T|N} >
решающая часть правила =>
исполнительная часть
</ *имя_правила*>

Правило имеет вид шаблона, переменные которого определены на множестве узлов дерева разбора. Для обозначения переменных используются буквы латинского алфавита с префиксом ' # '.

Пример 2. T-правило:

```
<ЧислРодРод ТИП = "T">
  #W1 Род #W2 & #W3 Род #W1 &
  ЧАСТЬРЕЧИ(#W1)= Числ &
   ЧАСТЬРЕЧИ(#W3)= Сущ &
   СЛОВО(#W3)=*_из &
   ЗНАЧАЩИЙ(#W2) != 0 =>
   ВСТАВИТЬ(#W3 Род #W2);
   УДАЛИТЬ(#W1 Род #W2)
</ЧислРодРод>
```

Правило применимо, например, к фрагменту текста (см. рис. 2) «… одна из двух галактических координат, измеренная …», представленного на дереве разбора набором триплетов

… «одна_из» *Какой* «измеренная»; *Род* «двух».

«двух» *Род* «**Галактический координата**» …

Опорным узлом для правила является W1 = «двух». При этом W2 = «**Галактический координата**», W3 = «одна_из», причем часть речи W1 – числительное, а значением W2 является один из определяемых терминов терминологического словаря.

Отметим, что синтаксический анализатор В.А. Тузова для словосочетания #W3 = «одна_из» вырабатывает ЧАСТЬРЕЧИ(#W3)= Сущ.

Результатом применения правила является набор триплетов

…«одна_из» *Какой* «измеренная»; *Род* «**Галактический координата**»; *Род* «двух» …

Высказывание о свойствах узлов контекста правила представляет собой стандартное И/ИЛИ логическое выражение, атомарными элементами которого являются предикаты, характеризующие тот или иной узел W. Использовались, например, следующие предикаты:

КЛАСС(#W) = имя класса универсальной онтологии, к которому относится лексема слова, сопоставленного узлу W дерева разбора;

ЧАСТЬРЕЧИ(#W)={Сущ, Прил, МС-С (местоименное сущ.), Глаг, МС_П (местоименное прил.), Прич, Деепр, Числ (количеств.), ЧислП (порядк.), Союз, Нареч};

ЗНАЧАЩИЙ(#W) != 0 – узлу W сопоставлен термин исследуемого терминологического словаря;

ЛИСТ(#W) != 0 – узел W – лист на дереве разбора в его текущем состоянии.

Функции, из которых набирается последовательность трансформаций в продукционной части правил:

УДАЛИТЬ(*триплет*) – удаление триплета, осуществляемое этой функцией не должно нарушать связность дерева разбора;

ВСТАВИТЬ(*триплет*) – субъектом вставляемого триплета должен быть один из узлов контекста правила;

ЗНАЧАЩИЙ(W) != 0 – узел W помечается как термин.

Приведенный список предикатов и функций не является исчерпывающим. Здесь приведено лишь то, что используется далее в примерах

Отметим, что каждый из узлов и триплетов дерева разбора в любой момент его обработки может находиться в одном из двух состояний, условно обозначенных как «обработан» или «не обработан». При этом имеются достаточно жесткие ограничения на использование триплетов в решающих частях правил:

1) Граф, представляемый набором триплетов в решающей части любого правила, должен быть связным.

2) Триплет, заданный первым в T-правиле, должен быть исходящим из опорного узла и находиться перед применением правила в состоянии «не обработан», а узел, к которому направлен этот триплет, – в состоянии «обработан». В контексте T-правила можно использовать лишь триплеты, принадлежащие пути из опорного узла к корню дерева разбора.

3) Опорный узел, заданный первым триплетом N-правила, должен находиться в состоянии «не обработан». В контексте правила могут быть использованы только исходящие из опорного узла триплеты, имеющие состояние «обработан», а также три-

плеты, принадлежащие пути из опорного узла к корню дерева разбора.

Два вида правил образуют пару строго упорядоченных списков: T-список и N-список. Применение правил из N-списка в отношении заданного опорного узла дерева разбора начинается строго после того, как исчерпаны возможности применения к этому опорному узлу правил из T-списка. Поиск подходящего правила в любом из списков всегда начинается с начала списка и осуществляется до первого подходящего правила. Продукционная часть найденного правила исполняется, и следующий поиск снова начинается с начала.

Пример 3. В этом примере мы покажем процедуру применения правил к дереву разбора, представленному на рис. 2.

А) Набор триплетов, представляющий исходное дерево разбора:

«**Галактический широта**» *ЭтоЕсть* «одна_из».
«одна_из» *Какой* «измеренная»; *Род* «двух».
«двух» *Род* «**Галактический координата**».
«измеренная» *Ото* «от_плоскости»; *кДат* «**к_объекту**».
«от_плоскости» *Род* «**галактика**».

Б) Список правил, применимых для данного набора триплетов:

```
<ЧислРодРод ТИП = "T">
 #W1 Род #W2 & #W3 Род #W1 &
 ЧАСТЬРЕЧИ(#W1)= Числ &
 ЧАСТЬРЕЧИ(#W3)= Сущ &
 СЛОВО(#W3)=*_из &
 ЗНАЧАЩИЙ(#W2) != 0 =>
 ВСТАВИТЬ(#W3 Род #W2);
 УДАЛИТЬ(#W1 Род #W2)
</ЧислРодРод>

<РодЧисл ТИП = "T">
 #W1 Род #W2 & ЗНАЧАЩИЙ(#W2) = 0 &
 КЛАСС(#W1) = Число =>
 УДАЛИТЬ(#W1 Род #W2)
</РодЧисл ТИП>

<РодНезн1 ТИП = "T">
 #W1 Род #W2 & ЧАСТЬРЕЧИ(#W1) = Сущ &
 КЛАСС(#W1) != Совокупность &
 ЗНАЧАЩИЙ(#W1) = 0 &
 ЧАСТЬРЕЧИ(#W2) = Сущ =>
 УДАЛИТЬ(#W1 Род #W2)
</РодНезн1>

<Ото1 ТИП = "T">
 #W1 Ото #W2 & (ЧАСТЬРЕЧИ(#W1) = Прич OR
 ЧАСТЬРЕЧИ(#W1) = Глаг)&
 ЗНАЧАЩИЙ(#W2) = 0 =>
УДАЛИТЬ(#W1 Ото #W2)
</Ото1>

<кДатЗнач ТИП = "T">
 #W1 кДат #W2 & #W3 Какой #W1 &
 ЧАСТЬРЕЧИ(#W1) = Прич &
 КЛАСС(#W1) = ОпределениеПараметров &
 ЗНАЧАЩИЙ(#W2) != 0 &
 ЧАСТЬРЕЧИ(#W3) = Сущ =>
 ВСТАВИТЬ(#W3 этоПараметр #W2);
 УДАЛИТЬ(#W1 кДат #W2)
</кДатЗнач>

<Какой ТИП = "T">
 #W1 Какой #W2 & (ЧАСТЬРЕЧИ(#W2) = Прил OR
 ЧАСТЬРЕЧИ(#W2) = Прич OR
 ЧАСТЬРЕЧИ(#W2) = МС-П) &
 ЗНАЧАЩИЙ(#W2) = 0 =>
 УДАЛИТЬ(#W1 Какой #W2)
</Какой>

<ЭтоОдна ТИП = "T">
 #W1 Род #W2 & #W3 ЭтоЕсть #W1 &
 ЗНАЧАЩИЙ(#W2) != 0 &
 КЛАСС(#W1) = Число & СЛОВО(#W1) = *_из =>
 ВСТАВИТЬ(#W3 subClassOf #W2);
 УДАЛИТЬ(#W1 Род #W2)
</ЭтоОдна>

<один_изПараметр ТИП = "T">
 #W1 этоПараметр #W2 & #W3 ЭтоЕсть #W1 &
 СЛОВО(#W1) = *_из &
ЗНАЧАЩИЙ(#W3) != 0 &
 НАЧАЩИЙ(#W2) != 0 =>
 ВСТАВИТЬ(#W3 этоПараметр #W2);
 УДАЛИТЬ(#W1 этоПараметр #W2)
</один_изПараметр>

<ЭтоЕстьНов1 ТИП = "N">
 #W1 ЭтоЕсть #W2 & ЗНАЧАЩИЙ(#W1) != 0 &
ЗНАЧАЩИЙ(#W2) = 0 &
КЛАСС(#W2) != ПАРАМЕТРЫ =>
ЗАМЕНИТЬ(#W1 ЭтоЕсть #W2, #W1 Это #W2);
СЛОВО(#W2)= Class;
ЛЕММА(#W2)= Class
</ЭтоЕстьНов1>
```

В) Пошаговый протокол преобразования исходного дерева разбора.

Шаг 1. Обрабатываемый триплет: «двух» *Род* «Галактический координата». Применяется правило "ЧислРодРод":

W1 = «двух»; W2 = «**Галактический координата**»; W3 = «одна_из».

Результат применения правила:

«**Галактический широта**» *ЭтоЕсть* «одна_из».
«одна_из» *Какой* «измеренная»; *Род* «двух»;
Род «**Галактический координата**».
«измеренная» *Ото* «от_плоскости»;
кДат «**к_объекту**».
«от_плоскости» Род «**галактика**».

(Связь '*Род* «**Галактический координата**»' переброшена с узла «двух» на узел «одна_из»).

Шаг 2. Обрабатываемый триплет «одна_из» *Род* «двух». Правило "РодЧисл":

W1 = «одна_из»; W2 = «двух»;
Результат:
«**Галактический широта»** *ЭтоЕсть* «одна_из».
«одна_из» *Какой* «измеренная»;
Род «**Галактический координата»**.
«измеренная» *Ото* «от_плоскости»;
кДат «**к_объекту**».
«от_плоскости» Род «**галактика**».
(Триплет «одна_из» *Род* «двух» удален.)

Шаг 3. Обрабатываемый триплет:
«от_плоскости» *Род* «**галактика**».
Правило "РодНезн1":
W1 = «от_плоскости»; W2 = «**галактика**».
Результат:
«**Галактический широта»** *ЭтоЕсть* «одна_из».
«одна_из» *Какой* «измеренная»;
Род «**Галактический координата»**.
«измеренная» *Ото* «от_плоскости»;
кДат «**к_объекту**».
(Удален триплет
«от_плоскости» *Род* «**галактика**»).

Примечание: безусловно правильным было бы не исключать триплет, а установить эквивалентность "плоскость галактиктики" = "галактическая плоскость" и, как следствие, получить зависимость галактической координаты от галактической плоскости, но в данном состоянии интерпретатора правил мы, к сожалению, не умеем выявлять эту достаточно распространенную перифразировку.

Шаг 4. Обрабатываемый триплет:
«измеренная» *Ото* «от_плоскости»
Результат применения правила "Ото1":
триплет удален.

Шаг 5. Обрабатываемый триплет:
«измеренная» *кДат* «**к_объекту**».
Правило "кДатЗнач":
W1= «измеренная»; W2= «**к_объекту**».
Результат:
«**Галактический широта»** *ЭтоЕсть* «одна_из».
«одна_из» *Какой* «измеренная»;
Род «**Галактический координата**»;
этоПараметр «**объект**» .
(Вставлен триплет
«одна_из» *этоПараметр* «**объект**», и удален триплет «измеренная» *кДат* «**к_объекту**»).

Шаг 6. Обрабатываемый триплет:
«одна_из» *Какой* «измеренная»
Результат применения правила "Какой" - триплет удален и:
«**Галактический широта**» *ЭтоЕсть* «одна_из».
«одна_из» Род «**Галактический координата**»;
этоПараметр «**объект**» .

Шаг 7. Обрабатываемый триплет:
«одна_из» *Род* «**Галактический координата**»
в контексте
«**Галактический широта**» *ЭтоЕсть* «одна_из»
Правило "ЭтоОдна":
W1= «одна_из»;
W2= **Галактический координата;**
W3= **Галактический широта.**
Результат:
«**Галактический широта**» ЭтоЕсть «одна_из»; subClassOf «**Галактический координата**».
«одна_из» этоПараметр «**объект**».
(Добавлен триплет:
«**Галактический широта**» subClassOf «**Галактический координата**» и удален
«одна_из» Род «**Галактический координата**»).

Шаг 8. Обрабатываемый триплет:
«одна_из» этоПараметр «**объект**» в контексте «**Галактический широта**» ЭтоЕсть #W1.
Правило "один_изПараметр":
W1= одна_из;
W2= «**объект**»
W3= «**Галактический широта**»**.**
Результат:
«**Галактический широта**» ЭтоЕсть «одна_из»; subClassOf «**Галактический координата**»; этоПараметр «**объект**».

Шаг 9. Правилом "ЭтоЕстьНов1" термин «**Галактический широта**» объявляется классом и удаляется триплет
«**Галактический широта**» ЭтоЕсть «одна_из».
Окончательный результат:
«**Галактический широта**» Это Class; subClassOf «**Галактическая координата**»; этоПараметр «**объект**».

Полученный результат легко может быть преобразован в запись на языке OWL и представляет собой фрагмент онтологии, извлекаемой из терминологического словаря:

```
ont:Галактический_широта
        rdf:type owl:Class;
        rdfs:subClassOf
        ont:Галактический_Координата;
        rdfs:subClassOf [owl:Restriction;
                owl:onProperty ont:_объект;
                owl:allValuesFrom ont:Объект].
```

Отметим, что в процессе преобразования исходное свойство *с_Параметром* было конкретизировано стандартным переименованием, использующим имя класса, полученного в качестве области значений этого свойства.

3 Основные результаты

3.1. Оценивая результат нашей работы, можно по-видимому говорить о достаточно четко просматриваемой технологии разработки терминологического словаря, если, конечно, речь идет о создании терминологического словаря одновременно с сопутствующей ему онтологией. Говоря точнее, мы в нашей работе исходили из предположения, что к моменту начала работы общий список терминов уже

составлен и для каждого из терминов подобрано вербальное определение. Какой-либо программной поддержкой этого этапа работы мы не занимались.

Далее следует рассматриваемый в докладе этап: автоматизированное выявление системы онтологических отношений между отобранными терминами и оформление выявленных отношений в виде формальной онтологии.

3.1.1. Работа начинается с предварительной обработки исходного терминологического словаря. Цели:

– пополнение исходного семантического словаря описаниями новой, используемой в исходных текстах лексики;

– получение списка терминов в виде отдельного информационного ресурса.

3.1.2. Далее ведется индивидуальная работа с каждым из определений. Итеративный цикл работы с конкретным определением включает:

3.1.2.1. Оценку результата семантико-синтаксического анализа определения по его визуализации. Подбор подходящей редакции определения в случае неудовлетворительного разбора.

3.1.2.2. Автоматическую трансформацию исходного дерева разбора по имеющейся системе правил (может быть хорошо поддержана программными средствами, разработан вариант такой программы). Оценку результата: при необходимости – либо редактирование исходного определения, либо пополнение системы правил.

3.1.2.3. Преобразование результата трансформации к представлению в виде онтологии (пример такого преобразования в докладе есть, см. пример 3). Подсоединение полученного фрагмента к общей онтологии терминологического словаря. Оценка результата: при необходимости – либо редактирование онтологии, либо пополнение системы правил.

3.2. Оценка объема и состава правил. Всего для анализа заданного терминологического словаря потребовалось 123 правила. Из них 51 правило было использовано два и более раз; 72 правила были использованы однократно.

Все правила, будучи ориентированными на задачу онтологического анализа терминологических словарей, вместе с тем имеют достаточно общий характер и не зависят от специфики конкретной предметной области (в нашем случае это задача межзвездного поглощения).

Необходимо отметить жесткую связь между системой правил и используемым семантико-синтаксическим анализатором. Связь эта проявляется как в общей номенклатуре имен связей, поставляемых анализатором, так и в свойственном анализатору методе конфигурирования дерева разбора.

3.3. Оценка объема необходимой лексикографической информации, поставляемой семантическим словарем.

При определении правил оказалась востребованной в полном объеме морфологическая информация и в меньшей степени – семантика лексических значений слов. Для рассматриваемого (и сильно ограниченного) терминологического словаря оказалась достаточной информация о принадлежности значения слова к одному из корневых или близких к корневым классам универсальной онтологии. Среди востребованных можно отметить классы физических объектов, абстрактных объектов, качеств, а также класс состояний и процессов.

К классу качеств, например, относятся значения общих слов – «свойство», «атрибут», «параметр», и более частные – «светимость», «яркость», «температура», «протяженность».

К классу состояний и процессов относятся, прежде всего, глаголы и отглагольные существительные. В этом классе выделяются подклассы слов, обозначающих вневременные состояния (стативы), например, «определять», «характеризовать», и агентивные процессы с явно выраженным объектом воздействия («измерять», «вычислять»).

К классу абстрактных объектов относятся, например, значения слов «скаляр», «число», «величина», а также такие словосочетания, как «разность между», «среднее значение».

Всего в определениях терминологического словаря использовано 237 различных слов. Указание класса потребовалось для 48 слов. Общее количество востребованных классов равно 17.

Литература

[1] Gomez F., Hull R., Segami C. Acquiring knowledge from encyclopedic texts. – http://acl.ldc.upenn.edu/A/A94/A94-1014.pdf.

[2] Brewster C. Techniques for automated taxonomy building: Towards ontologies for knowledge management. – http://eprints.aktors.org/129/01/ BrewsterCLUK02. pdf.

[3] Рубашкин В.Ш., Капустин В.А. Использование определений терминов в энциклопедических словарях для автоматизированного пополнения онтологий // В сб.: «Языковая инженерия: в поисках смыслов». Доклады семинара «Лингвистические информационные технологии в Интернете»: XI Всерос. объединенная конф. «Интернет и современное общество». – СПб., 2008. – С. 32-39.

[4] Лукашевич Н.В., Салий А.Д., Добров Б.В. Использование компьютерных технологий для экспертизы терминологического словаря в области государственного финансового контроля // Компьютерная лингвистика и интеллектуальные технологии: Труды межд. конф. Диалог'2005 (Звенигород, 1 – 6 июня 2005 г.).

[5] Брюхов Д.О., Вовченко А.Е., Захаров В.Н., Желенкова О.П., Калиниченко Л.А., Мартынов Д.О., Скворцов Н.А., Ступников С.А. Архитектура промежуточного слоя предметных посредников для решения задач над множеством интегрируемых неоднородных распределенных информационных ресурсов в гибридной грид-инфраструктуре виртуальных обсерваторий.

//Информатика и её применения. – 2008. – Т. 2, Вып. 1. – С. 2-34.
[6] Скворцов Н.А., Ступников С.А. Использование онтологии верхнего уровня для отображения информационных моделей // Труды 10-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008. – С. 122-127.
[7] OWL Web Ontology Language Guide. W3C Recommendation 10 February 2004. – http://www.w3.org/TR/2004/REC-owl-guide-2004021.
[8] Тузов В.А. Компьютерная семантика русского языка. – СПб.: Изд-во СПб ГУ, 2004. – 400 с.
[9] Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., Schneider L. DOLCE: a Descriptive Ontology for Linguistic and Cognitive Engineering // DOLCE documentation. – http://www.loa-cnr.it/DOLCE.html.
[10] RDF Primer. W3C Recommendation 10 February 2004. – http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

Automation of process of extraction of the ontological information from verbal terminological dictionaries (on the example of the terminological dictionary of the problem of interstellar extinction)

K.K. Boyarsky, E.A. Kanevsky, G.V. Lezin
L.A. Kalinichenko, N.A. Skvortsov

In the article the problem of construction of ontological model of a subject domain under its specification set by the terminological dictionary is considered. Algorithms of extraction of the ontological information from the terminological dictionary are set by a collection of the production rules applied to result of the semantic-syntactical analysis of definitions of the dictionary. The program of interpretation of such rules is developed and experiment on working out of rules and their application for the small highly specialized dictionary is made. In the article the preliminary analysis of results of experiment is described.