Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках

Андреев А.М., Березкин Д.В., Симаков К.В.

НПЦ «ИНТЕЛТЕК ПЛЮС»
info@inteltec.ru

Аннотация

В статье предложен подход к решению задачи поиска противоречий в правовых текстах. Описана структура модели предметной области и онтологии. Представлены примеры противоречий, предложена их формальная модель и методы выявления, опирающиеся на структуру модели и онтологии. Рассмотрена структура ЭБ «Мониторинг правового пространства и правоприменительной практики в СФ РФ».

Введение

Термин «противоречие» охватывает обширное поле проблем, решаемых в таких областях науки как логика, лингвистика, юриспруденция, психология, этика и др. Многие аспекты противоречий (морально-этические, психологические) не поддаются строгой формализации, что делает затруднительным их автоматическое выявление с применением вычислительной техники. Тем не менее, логическая сторона противоречий, разработанная еще в античности (Аристотель), поддается строгой формализации, что в принципе дает возможность ее реализации методами искусственного интеллекта.

Обобщенно, в данной работе рассматриваются противоречия, обнаруживаемые в естественно-языковых текстах из правовой области, квалифицируемые как разного рода несоответствия ее модели (МПО) и онтологии (ОПО).

Предложенные структуры МПО и ОПО основаны на разработанных ранее моделях и методах семантического анализа текстов [1], применяемых в системе понимания текстов (СПТ) [2]. Основная цель данной работы – разработка системы поиска противоречий в тестах правовых документов, использующую для этого модель и онтологию.

Также необходимо сделать следующее допущение. Приводя в пример действующее законодательство РФ, можно заметить, что вполне правомерным является наличие взаимоисключающих законов, взятых из разных его областей. Поэтому, в предлагаемом подходе считается, что модель и онтология предметной области строятся для узкой сферы законодательства, внутри которого противоречий быть не должно.

1. Примеры противоречий в законодательстве

Прежде, чем разрабатывать теорию выявления противоречий, были исследованы реальные случаи выявления противоречий в законотворческой практике. Приводимый перечень примеров отнюдь не является полным, и служит иллюстрацией рассматриваемой в данной работе проблемы. Более того, технология построения системы выявления противоречий должна в качестве одного из первых этапов включать подобный анализ имеющихся примеров, выполняемый экспертом-аналитиком. Такой анализ позволит позднее сформулировать формальные правила, обнаруживающие противоречия в новых текстах. Иными словами, прежде чем закладывать в систему правила выявления противоречий необходимо определить их классификацию по результатам анализа естественно-языкового материала.

1.1. Противоречия в понятиях

Включение в текст документа (раздела законодательства) понятия и его прямого отрицания, выраженного в явной и неявной форме.

Пример:

В Федеральном законе “О реструктуризации кредиторской задолженности юридических лиц по налогам и сборам в бюджеты всех уровней”, принятом Государственной Думой, но отклоненном Советом Федерации, дано понятие задолженности по финансовым санкциям как задолженности по пеням и штрафам. В части первой Налогового кодекса РФ применяется понятие “налоговой санкции” как меры ответственности за совершение налогового правонарушения, устанавливаемой и применяемой в виде денежных взысканий (штрафов). Пеня, по своей природе не являющаяся налоговой или финансовой санкцией, рассматривается в Налоговом кодексе как один из способов обеспечения исполнения обязанностей по уплате налогов и сборов.

Введение в тексте документа нового понятия, совпадающего по смыслу с определенным ранее в законодательстве понятием.

Пример:

В указанном выше проекте Федерального закона, дано определение кредиторской задолженности по налогам и сборам в бюджеты всех уровней. Она определяется как задолженности по обязательным платежам, не внесенным организацией в бюджеты в установленные законодательством РФ сроки, сложившейся по состоянию на определенную дату. В части первой Налогового кодекса РФ такая задолженность определена понятием "недоимка".

Использование неопределенных понятий.

Пример:

Цепочка отсылок к другим законам или подзаконным актам типа: «как это установлено в Законе таком-то», из этого закона ссылка на то, что определяемое должно определяться Правительством РФ, в Постановлении Правительства РФ поручение какому-то ведомству разработать инструкцию, которой пока нет.

1.2. Противоречия в предикатах

Неверный предикат.

Пример:

Вместо фразы "Президент РФ предлагает кандидатуру Генерального прокурора РФ", не правильным будет предложение: "Президент РФ назначает Генерального прокурора РФ".

Президент РФ подписывает законы, издает указы, … , изменяет ставки налогов. Последний предикат противоречит законодательству.

Нарушение порядка применения предикатов.

Пример:

В Гражданском кодексе РФ установлен судебный и внесудебный порядок изъятия заложенного имущества, причем последний применяется лишь в отдельных случаях. Федеральный закон "О реструктуризации кредиторской задолженности юридических лиц по налогам и сборам в бюджеты всех уровней", принятый Государственной Думой, но отклоненный Советом Федерации изменял этот порядок и устанавливал исключительно внесудебный порядок изъятия заложенного имущества, что противоречило Гражданскому кодексу РФ и Федеральному закону "Об исполнительном производстве".

1.3. Противоречия иного типа

К противоречиям иного типа, относится, противоречие между отдельными областями права, например, противоречия между гражданским и публичным правом. Так в Гражданском кодексе РФ (Статья 855) и в и налоговом законодательстве использовался разный приоритет платежей. Это противоречие было в свое время устранено решением Конституционного суда, а затем, и внесением соответствующих изменений в Закон РФ "Об основах налоговой системы в РФ".

Другими примерами являются: апелляции к нормам международного права, морали, здравого смысла, а так же несоответствия стилистического характера. К последнему относятся такие ситуации, когда содержание преамбулы закона или его статей не соответствует названию, названия отдельных статей или глав не отражают их содержание, входящие в название закона ключевые понятия (термины) могут нигде в тексте не использоваться и т.д.

Для указанных видов противоречий может оказаться затруднительным построение их математических моделей, тем не менее, успешная формализация противоречий данного типа обеспечивает их выявление наравне с остальными.

2. Модель и онтология

Зачастую понятия модели и онтологии предметной области используются, как синонимы [18]. В данной работе они двусторонне отражают семантику предметной области, аналогично [8], где предметная область представлена тезаурусом и онтологией. Тезаурус используется в [8] как инструмент семантического анализа на общелингвистическом уровне, онтология применяется для семантического анализа на уровне предметной области.

В данной работе МПО разрешает задачи лингвистического и семантического анализа текста на естественном языке (ЕЯ) с построением его семантического представления, в то время как ОПО служит для решения частных задач, таких как выявление противоречий.

2.1. Структура модели предметной области

Необходимость введения модели предметной области отмечается во многих источниках. В частности в [5] вводится понятие концептуальной модели мира, включающей в себя описания базовых понятий, организованных в родовидовые деревья, и совокупность связей между ними. Аналогия прослеживается и в работе [6], где концептуальная модель включает в себя описание объектов, понятий и отношений действительности. В обеих работах кроме концептуальных моделей определяется привязка введенных описаний к языковым средствам, выражающим эти описания в естественных текстах. Так в [5] вводится идеографический словарь предметной области, лексически наполняющий концептуальную модель, а в [6] такая привязка определяется как фрагмент базы знаний, в котором указаны соответствия между языковыми единицами и элементами концептуальной модели.

В данной работе модель предметной области складывается из таксономии свойств и библиотеки семантических отношений. В таком виде модель используется для непосредственного семантического анализа текстов (см. [1]), без дополнительного привлечения онтологии.

Таксономия свойств используются для смыслового разделения лексики предметной области [9]. Следует различать свойства и их значения. Свойства указывают точки смыслового дробления лексики, а их значения определяют области, получаемые в результате такого дробления. Формально таксономия

свойств, как алгебраическая система, обозначается следующим образом

$$M_1 =< D, B, R_1, R_2 > \qquad (1)$$

где D – множество свойств, B – множество их значений, R_1 – отношение на D×B, R_2 – отношение на B×D. Отношения R_1 и R_2 наделены следующими особенностями:

1. $\forall d_i \in D\ \exists B_i \subset B : \forall b \in B_i \rightarrow (d_i, b) \in R_1 \equiv d_i R_1 b \wedge |B| > 1$, то есть отношение R_1 определяет принадлежность значений из множества B конкретным свойствам множества D, причем каждому свойству соответствует не менее двух значений.
2. $\forall b \in B\ \exists! d \in D : dR_1 b$, согласно этой особенности во множестве B не существует значений, несвязанных отношением R_1 ни с одним свойством, кроме того, каждое значение связано отношением R_1 только с одним свойством.
3. $\exists! d_0 \in D \rightarrow \neg\exists b \in B : bR_2 d_0$, согласно этой особенности во множестве D существует единственный корневой элемент, несвязанный ни с одним значением отношением R_2.
4. $\forall d \neq d_0 \in D\ \exists! b \in B : bR_2 d$, согласно этой особенности всякое свойство, не являющееся корневым, связано отношением R_2 с единственным элементом из B.

   $\forall d, b : (d, b) \in R_1 \rightarrow (b, d) \notin R_2$
5. $\forall b, d : (b, d) \in R_2 \rightarrow (d, b) \notin R_1$, согласно этой особенности любая пара (свойство,значение) не может одновременно находиться в отношении R_1 и R_2.

На рис. 1 дана графическая интерпретация модели M_1(см. также в [1]), вытекающая из ее пяти свойств, которые позволяют вывести (доказать теорему существования) отношение семантической совместимости R_3 на множестве B.

Отношение R_3 для любой пары (bi,bj) отвечает на вопрос о смысловой совместимости этих значений. Отношение совместимости R_3 обладает следующими свойствами:

1. Рефлексивность: $\forall b \in B \rightarrow bR_3 b$
2. Симметричность: $\forall b_i, b_j \in B : b_i R_3 b_j \rightarrow b_j R_3 b_i$

В качестве примеров, согласно рис. 1, могут быть приведены следующие пары, принадлежащие R_3: (b_9,b_5), (b_9,b_8), (b_6,b_3), (b_6,b_8) и т.д. И наоборот пары (b_9,b_{10}), (b_9,b_{11}), (b_7,b_4) и т.д. не принадлежат R_3. Используя R_3 значения можно объединять в наборы по следующему правилу (2)

$$\forall t_i \in T \rightarrow t_i = B_i \subseteq B : \forall b_k, b_j \in B_i \rightarrow b_i R_3 b_j \qquad (2)$$

Где T – множество всех допустимых в модели M_1 наборов значений, B_i – счетное подмножество множества B модели M_1. Согласно (2) всякий набор $t_i \in T$ содержит элементы из $B_i \subseteq B$ такие, что любая их пара (b_i,b_j) находится в отношении совместимости R_3 модели M_1. Далее эти наборы значений будут участвовать в качестве семантической компоненты понятий, извлекаемых из текста.

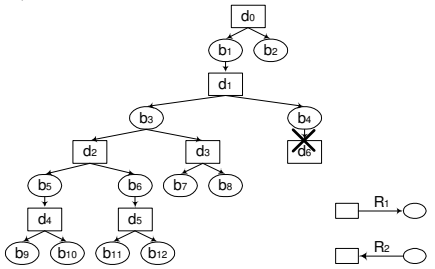

Рис. 1. Иллюстрация таксономии свойств.

Для большей наглядности дадим естественно языковую интерпретацию таксономии. Она задает состав и взаимосвязи свойств понятий предметной области. Любое понятие складывается из имени и набора свойств. Имя представлено словом или словосочетанием естественного языка, которое употребляется в текстах для указания на данный объект. Свойства же отражают семантику (смысл) данного объекта, поскольку одного имени для обозначения смысла объекта не достаточно в силу полисемии естественного языка. По этой же причине, предметная область описывается не явной иерархией понятий (как, например, в [20]), а таксономией их свойств. Сами же понятия извлекаются в процессе семантического анализа текстов, при этом им приписываются свойства, заложенные в таксономии. Пример таксономии приведен на рис. 2.

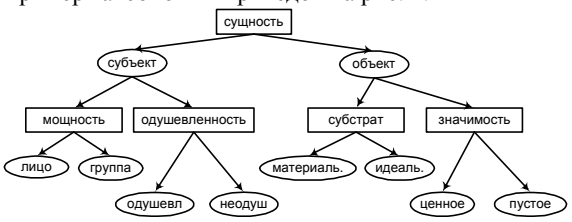

Рис. 2. Пример таксономии свойств.

Наличие отношения семантической совместимости между значениями свойств позволяет утверждать (см. рис. 2), что «лицо» и «одушевленное» могут быть объединены в набор значений, аналогично «лицо» и «неодушевленное», поскольку оба эти значения могут быть присущи понятиям предметной области. В то же время значения свойств «лицо» и «материальное» в данной модели не являются совместимыми, поэтому появление в процессе семантического анализа набора {«лицо», «материальное»} говорит о противоречии анализируемого текста самой модели предметной области.

Библиотека семантических отношений вводит описания связей между возможными объектами предметной области. При этом сами объекты в модели не определяются, а возможные участники описывается наборами значений свойств (определяемых согласно (2)), значимыми для выполняемой роли в данном отношении. Библиотека семантических отношений определяется как алгебраическая система вида $M_2 =< L, N, T, F, R_4, R_5, R_6 >$ (3)

Где L – множество семантических отношений, определенных в предметной области, N –

подмножество натуральных чисел, T – определенное в (2) множество наборов значений свойств, F – множество формул V – языка, подробное описание которого дано в [3]. R_4 – отношение на L×N, R_5 – отношение на R_4×T, R_6 – отношение на L×F.

При этом каждая пара $(l,n)\in R_4$ определяет n-ого возможного участника отношения l, где n используется для уникальной идентификации участников внутри отношения l. Каждая пара $((l,n),t)\in R_5$ определяет набор значений свойств, характерный для n-ого участника отношения l. Существенно то, что сам участник не определен, известен только t - характерный для участника набор значений. Каждая пара $(l,f)\in R_6$ определяет характерную для отношения l синтаксическую конструкцию естественного языка, описываемую частично определенной V – формулой f. Каждой переменной формулы f соответствует некоторый n-ый участник отношения l. (Аналогичным образом описывается синтаксис предложений в [11] с использованием категориальной грамматики; в [12] предложение представляется в виде суперпозиции функций {fi}; в [14],[16] для этих целей используются фразовые образцы «phrasal patterns»).

Таким образом, отношения представляют собой предикаты типа l(t1,..,tn), где t1,…,tn наборы значений свойств из описанной таксономии, характерные для соответствующих участников отношения. Имена участников не указываются, поскольку одно и то же отношение может быть использовано для взаимодействия разных группа объектов. Для связи с естественным языком отношение должно содержать естественно языковые шаблоны его применения в текстах, задаваемые с помощью формул V – языка [3]. Пример отношения приведен на рис. 3. В данном примере дана запись библиотеки отношений, описывающая отношение l1, с которым связываются два участника.

```
l1=отношение с двумя участниками:
t1={субъект}
t2={ценное,материальное}
V1(X,V2(C,Y))
C=обязан платить
X-отражает первого участника
Y-отражает второго участника
V1-согласование подлжеащего и сказуемого
V2-согласование сказуемого и дополнения
```

Рис. 3. Пример отношения.

С первым участником ассоциируется набор значений свойств, состоящий из одного элемента {субъект}, со вторым участником – набор {ценное,материальное}. Также с отношением связан V - шаблон естественно языковой конструкции V1(X,V2(C,Y)), который переменной X – ставит в соответствие первого участника отношения, а переменной Y – второго участника. Константа C задает ключевое словосочетание «обязан платит», по которому будет отыскиваться шаблон в библиотеке. Операционная константа V1 регламентирует синтаксическую роль первого участника отношения в предложениях как подлежащего, операционная константа V2 требует, чтобы второй участник отношения играл роль дополнения в анализируемых предложениях, связанного со сказуемым C (обязан платить) связкой V2(C,Y).

В процессе анализа текста в нем ищутся фрагменты, соответствующие шаблонам, заложенным в библиотеке отношений. Из фрагментов текстов, удовлетворяющих некоторому шаблону отношения l(t1,..,tn), извлекаются имена объектов и им приписываются наборы значений t1,…,tn соответствующих участников отношения. Приписав именам объектов, наборы значений соответствующих участников, формируются понятия вида

$$o = (V^{a(...)}; t) \quad (4)$$

Где $V^{a(...)}$ - терм V – языка, выделенный во входной формуле, t - набор значений модели M_1. Терм отражает лингвистическую составляющую понятия (множество возможных лексем или устойчивых словосочетаний, принятых для выражения данного понятия средствами естественного языка); второй – отражает смысл понятия. Между выделенными понятиями устанавливаются связи – отношения из модели M_2, на основе которых и было выполнено извлечение. Таким образом, формируется сетевая структура, называемая далее семантической сетью [13].

2.2. Онтология предметной области

Онтология задается в виде базовых понятий и отношений между ними. В отличие от модели, в которой заложены наиболее общие закономерности: отношения, их лингвистические репрезентации в виде конструкций V – языка, наборы значений участников этих отношений, онтология вводит базовые понятия и установленные между ними отношения [19]. Иными словами онтология представляется в виде семантической сети, так же как и описание смысла отдельного текстового документа. В таком виде, онтология выступает, как объединение всех семантических представлений текстов из корпуса предметной области в единую сеть.

В связи с этим, получение онтологии может быть выполнено автоматически, с минимальным участием эксперта, если представится возможность описать онтологию в виде набора тестов на естественном языке.

Формирование полного семантического представления текста выполняет средствами глобального семантического анализа. Методы формирования семантического представления текста изложены в [1], здесь ограничимся лишь его формальным определением. Семантическое представление текста имеет следующую структуру

$$M_t = < O_t, A_t, L_t, R_t^1, R_t^2 > \quad (5)$$

Где O_t – множество выделенных в тексте понятий вида (4), A_t – множество ребер, связывающих понятия O_t, $L_t \subset L$ – множество выявленных в тексте семантических отношений модели M_2, используемых как метки ребер A_t. R^1_t– отношение инцидентности на $O_t \times A_t \times N$, где N – подмножество идентификаторов участников отношений модели M_2. R^2_t – отношение инцидентности на $A_t \times L_t$. Таким образом, семантиче-

ская сеть текста состоит из понятий O_t, связанных ребрами A_t. Семантика ребер A_t определяется приписанными им отношениями L_t, роли понятий в отношениях L_t определяются номерами участников N_t. Поскольку онтология рассматривается как семантическое представление объединения корпуса текстов предметной области, то ее структура будет идентичной (5), но для ее обозначения и обозначения ее элементов используется запись (6)

$$M_o =< O_o, A_o, L_o, R_o^1, R_o^2 > \qquad (6)$$

Пример фрагмента онтологии, получаемой после семантического анализа предложения: «Субъект права обязан платить налоги и сборы» приведен на рис. 4, где l1 – отношение, приведенное на рис. 3. Такое понимание, онтологии часто используется в методах обработки естественно-языковых текстов ([16], [19], [20]), отличие заключается в том, что, как правило, набор знаний, закладываемых в интеллектуальную систему, не ограничивается одной лишь онтологией.

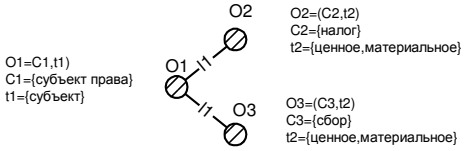

Рис. 4. Фрагмент онтологии.

Так в [16] и [21] используется иерархия понятий, в [17] используются продукции, в [15] используются дополнительные грамматики, характеризующие особенности предметной области. В данной работе предметная область описывается моделью и онтологией. Кроме того, вводятся модели противоречий, которые по своей сути являются продукциями, записанными на языке предикатов первого порядка, оперирующих с элементами семантической структуры онтологии и анализируемого текста.

3. Анализ текстов и получение онтологии

Полная схема функционирования анализатора, формирующего семантическое представление текста приведено в [1] и [2].Здесь же ограничимся общей структурной схемой.

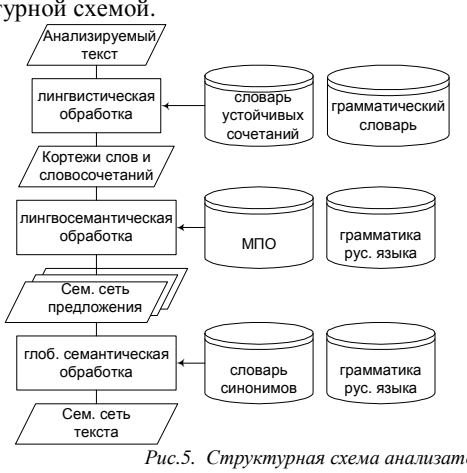

Рис.5. Структурная схема анализатора.

В данной работе приводятся некоторые особенности блока лингвосемантической обработки.

3.1. Лингвосемантический анализатор

На рис. 6 приведена функциональная схема лингвосемантического анализа. Данный анализ является локальным [7], поскольку формирует семантическую сеть для отдельных предложений. Существует так же глобальный семантический анализ, задача которого объединить сети отдельных предложений в более крупные фрагменты [15], отражающие семантику всего текста. В данной работе механизмы глобального анализа рассматриваться не будут, отметим лишь, что основными методами являются отождествление семантически совместимых понятий из разных фрагментов текста на основе поиска местоимений, синонимов и слов, связанных принадлежностью к одному классу.

Рис. 6 иллюстрирует шаги, предпринимаемые при формировании понятий и связей между ними для отдельных предложений текста, описанные ранее, поэтому здесь заострим внимание лишь на некоторых особенностях предлагаемого механизма.

На вход анализатора поступают подготовленные кортежи слов/словосочетаний, т.к. каждому предложению ставится в соответствие кортеж. Кортеж проходит ряд традиционных стадий [2] предварительной сугубо лингвистической обработки, не показанной на рис. 6. К таким стадиям относятся: выделение слов, сокращений и обозначений, поиск устойчивых словосочетаний, определение грамматических значений слов: часть речи, род, число, падеж и др.

Далее, как правило, должна следовать стадия синтаксического разбора. Но в силу существующих проблем, связанных с большим числом переборов, выполняемых современными синтаксическим анализаторами, в данном подходе предлагается объединить синтаксический и семантический анализ в единую стадию лингвосемантической обработки.

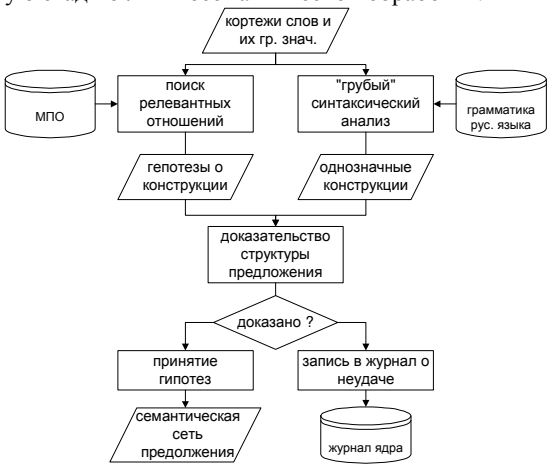

Рис. 6. Лингвосемантический анализ.

Задача лингвосемантической обработки – выявить понятия и семантические отношения между ними внутри отдельных предложений. Это выполняется одновременно с определением полной синтаксической структуры предложений (поэтому этап назван лингвосемантическим). Исследования мно-

гих разработок [15] косвенно показывают целесообразность именно такого подхода.

Таким образом, параллельно выполняются два вида анализа: поиск релевантных отношений и «грубый» синтаксический анализ.

Поиск релевантных отношений производится в МПО. Данный вид поиска по ключевым словам отыскивает в МПО характерные для искомых слов семантические отношения. Поскольку, семантическое отношение имеет набор ключевых слов (преимущественно глаголы), отражающие это отношение в текстах предметной области. Кроме того, отношения имеют описания характерных синтаксических конструкций в виде устоявшихся синтаксических шаблонов, описанных на V - языке. Таким образом, по ключевому слову в МПО отыскивается отношение, строится гипотеза о вероятных участниках этого отношения, а так же о синтаксической структуре фрагмента предложения, из которого взято слово. Вероятными участниками отношения объявляются слова, подобранные из этого фрагмента.

Параллельно с поиском релевантных отношений выполняется грубый синтаксический анализ. Данный вид анализа выполняется сугубо на основе грамматических значений слов независимо от контекста. Задача такого анализа сводится к установлению элементарных зависимостей между словами, таких как образование именной и глагольной группы, свертка однородных членов, причастных и деепричастных оборотов. В случае относительно простых и недвусмысленных предложений, данный вид анализа способен получить структуру всего предложения. В этом случае дальнейший лингвосемантический анализ сведется к подтверждению подходящих гипотез, сформулированных поиском релевантных отношений. На данном этапе анализа используется словарь с грамматическими правилами русского языка. Словарь должен содержать как общеупотребительные правила конструирования предложений, так и характерные для предметной области правила. Примером общеупотребительного правила является согласование в падеже, числе и роде существительного с прилагательным. Правилом, характерным для предметной области, является порядок слов в группе, например, прилагательное обычно предшествует существительному («гражданское право», но не «право гражданское»). Результатом работы грубого анализатора являются однозначно сформированные синтаксические конструкции фрагментов предложений, описываемые средствами V – языка.

Результат работы модуля поиска релевантных отношений и грубого синтаксического анализатора обрабатываются на этапе доказательства структуры предложения. Это необходимо для отбора «правильных» гипотез, выданных при поиске релевантных отношений, оставив именно «правильные» гипотезы, автоматически оставляются только те найденные семантические отношения и их участники, которые соответствуют данным гипотезам. В этом и заключатся неразрывная взаимосвязь лингвистической и семантической обработки данного этапа: с одной стороны строится полная структура предложения, а с другой стороны окончательно формируются семантические отношения и их участники, найденные в предложении. Суть доказательства структуры предложения сводится к выбору подходящих гипотез для стыковки с результатом грубого синтаксического анализа, а так же для дополнения этого результата, в случае, если грубый анализатор не построил цельную структуру предложения. Если доказательство выполнено успешно, т.е. выбран набор гипотез, позволяющих связать все слова предложения, то результатом работы анализатора являются связки понятий и отношений, соответствующих принятым гипотезам. Так же возможен вариант, когда разбор осуществить не удалось, это может быть вызвано, во-первых, неполнотой лингвистического обеспечения или МПО, во-вторых, не характерностью предложения для данной предметной области. Чтобы понять, какой из вариантов задействован, система записывает предложение в журнал ядра, который позднее может быть проанализирован экспертом. Так же возможен вариант, когда несколько конкурирующих наборов гипотез приводят к положительному результату, т.е. в нескольких случаях удается доказать несколько структур предложения, при этом каждой структуре ставится в соответствие несколько вариантов семантических сетей. Это явление называется синтаксической омонимией, задача ее разрешения выполняется на этапе глобальной семантической обработки (см. [4]).

3.2. Получение онтологии

Как отмечалось, онтологию планируется получить автоматически, при наличии репрезентативного корпуса текстов. Подобранный корпус подлежит семантической обработке согласно описанной выше функциональной схеме (рис. 5). В качестве примера рассмотрим формирование семантической сети представленной на рис. 5 для предложения «Каждый обязан платить законно установленные налоги и сборы» (ст. 57, Конституция РФ). На рис. 7 отображены все шаги, взятые из функциональной схемы для данного предложения. После лингвистической обработки предложение будет представлено кортежем слов/словосочетаний C1,C,C2,C3,C4,C5.

Для преобразования данного предложения традиционной грамматики русского языка может оказаться не достаточным, поскольку слово «каждый» не является существительным, но, тем не менее, выступает в роли подлежащего, так же с точки зрения смысла предложения это слово именует некоторый объект предметной области, а значит, это слово должно лечь в основу понятия. По этой причине «грубый» синтаксический анализ, используя грамматику русского языка, сможет разобрать только фрагмент предложения «законно установленные налоги и сборы» (кортеж C2,C3,C4,C5), что отражено в првой ветви рис. 6. Остальная часть предложения должна быть разобрана с применением МПО. Для этого в библиотеке семантических отношений

по ключевому словосочетанию «обязан платить» отыскивается отношение l1. Второму участнику (Y) синтаксического шаблона этого отношения V1(X,V2(C,Y)) удовлетворяет словосочетание «законно установленные налоги и сборы», представленное термом V3(C2,V4(C3,C4,C5)) в результате «грубого» синтаксического анализа. Поэтому в качестве первого участника (X) подставляется слово C1=каждый.

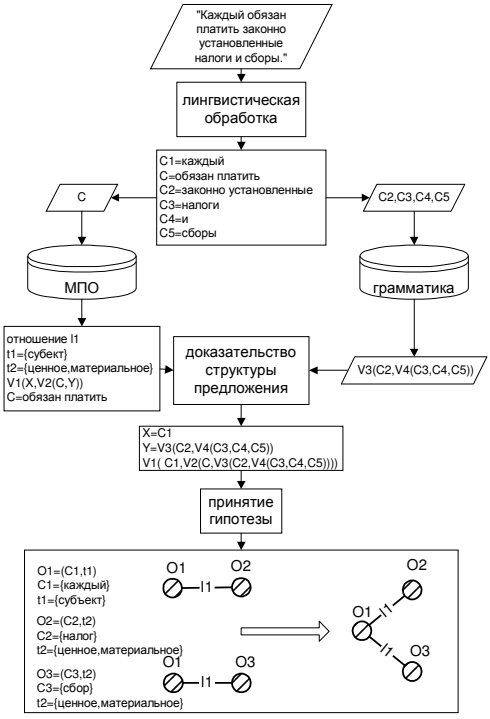

Рис. 7. Получение онтологии.

Поскольку требования шаблона удовлетворены, анализатор принимает гипотезу о наличие в анализируемом предложении отношения l1, образует три понятия O1, O2, O3 и связывает их этим отношением. Понятие O1 выражено словом «каждый», которому приписан набор значений свойств из таксономии свойств на рис. 2 {субъект}, понятия O2 и O3 выражены словами «налог» и «сбор» соответственно, им приписаны одинаковые наборы значений {ценное, материальное}.

4. Выявление противоречий в текстах

Далее рассмотрены модели выявления противоречий в анализируемом тексте с использованием терминологии и обозначений введенных ранее. Предполагается, что текст подвергнут семантическому анализу, так что получено его семантическое представление вида (5). Так же полагается, что кроме моделей M_1 и M_2 используемых анализатор, задана онтология предметной области M_o вида (6).

Рассматриваемые примеры записаны в виде высказываний на языке предикатов первого порядка.

Для упрощения записи высказываний далее используются следующие обозначения: $o = (X,t) \in M_o$ – некоторое понятие o принадлежит онтологии, т.е. содержится во множестве O_o из (6). Такие понятия также называются базовыми. $L_o(o)$ – подмножество множества L_o отношений из (6), размечающих ребра из A_o, инцидентные понятию o онтологии. $L_t(o)$ – подмножество множества L_t отношений из (5), размечающих ребра из A_t, инцидентные понятию o семантического представления текста. $O_o(o)$ – подмножество множества O_o понятий из (6), смежных с понятием o онтологии. $O_t(o)$ – подмножество множества O_t понятий из (5), смежных с понятием o семантического представления текста. $L_t(O_i)$ – подмножество множества L_t отношений из (5), размечающих ребра из A_t, инцидентные понятиям некоторого подмножества $O_i \subseteq O_t$ семантического представления текста. $L_o(O_i)$ – подмножество множества L_o отношений из (6), размечающих ребра из A_o, инцидентные понятиям некоторого подмножества $O_i \subseteq O_o$ онтологии. $L_t(o_i) \subseteq L_o(o)$ – отношения, в которые вступает понятие o_i в тексте, предусмотрены в онтологии для понятия o, при этом понятие o_i в тексте выступает в тех же ролях, что и базовое понятие o в онтологии, т.е.

$\forall l_t \in L_t(o_i) : l_t R_t^2 a_t \wedge (o_i, a_t, n_t) \in R_t^1 :$

$\exists l_o \in L_o(o) : l_o R_o^2 a_o \wedge (o, a_o, n_o) \in R_o^1 \wedge n_t = n_o$

$O_t(o_i) \subseteq O_o(o)$ – аналогично предыдущему обозначению: смежные понятия для o_i в тексте совпадают с предусмотренными смежными понятиями для o в онтологии, при этом роли понятий из $O_t(o_i)$ совпадают с ролями базовых понятий из $O_o(o)$, т.е.

$\forall o_t \in O_t(o_i) : (o_t, a_t, n_t') \in R_t^1$

$\exists o_o \in O_o(o) : (o_o, a_o, n_o') \in R_o^1 \wedge n_t' = n_o' \wedge o_t = o_o$

4.1. Противоречия в понятиях

4.1.1. Отрицание понятия

Имеется семантическое представление M_t анализируемого текста вида (5). Некоторое понятие $o_i = (X_i, t_i) \in M_t$ отрицает существующее базовое понятие онтологии M_o, если истинно одно из следующих утверждений.

1. $\exists o = (X,t) \in M_o : X = X_i \wedge t_i \cup t \notin T$, понятие o_i выражено словом или словосочетанием X, которое уже зарезервировано в онтологии для отражения базового понятия o. При этом понятия o и o_i имеют несовместимую семантику. Несовместимость выявляется в том случае, когда одно или более значений свойств из набора понятия o не находится в отношении R_3 семантической совместимости с каким-либо значением свойства из набора понятия o_i.
2. $\exists o = (X,t) \in M_o : X = X_i \wedge t_i \cup t \in T \wedge L_t(o_i) \not\subset L_o(o)$, понятие o_i выражено словом или словосочетанием X, которое уже зарезервировано в онтологии для отражения понятия o. При этом o имеет совместимую семантику с o_i, но в анализируемом тексте понятие o_i встречается в отношениях, не предусмотренных в онтологии для понятия o.

3. $\exists o=(X,t)\in M_o : X=X_i \wedge t_i \cup t\in T \wedge O_t(o_i)\not\subset O_o(o)$, аналог предыдущего случая, но в анализируемом тексте o_i вступает в отношения с такими понятиями, которые не предусмотрены онтологией для o.

4. $\exists o=(X,t)\in M_o : X=X_i \wedge t_i \cup t\in T \wedge L_t(o_i)\not\subset L_o(o) \wedge O_t(o_i)\not\subset O_o(o)$, данный случай представляет собой сочетание логическим И предыдущих двух вариантов.

4.1.2. Дублирование понятия

Имеется семантическое представление M_t анализируемого текста вида (5). Истинность одного из следующих утверждений выявляет факт дублирования понятием $o_i=(X_i,t_i)\in M_t$ некоторого базового понятия онтологии M_o.

1. $\exists o=(X,t)\in M_o : X\neq X_i \wedge t_i \subseteq t$, понятие o_i выражено словом или словосочетанием X, которое еще не зарезервировано в онтологии. При этом в онтологии обнаружено базовое понятие o такое, что смысловое содержание проверяемого понятия o_i является частью смыслового содержания базового понятия o.

2. $\exists o=(X,t)\in M_o : X\neq X_i \wedge t_i \subseteq t \wedge L_t(o_i)\subseteq L_o(o)$, менее строгое правило, выявляющее противоречие в случае, если проверяемое понятие является частью смыслового содержания базового понятия, а так же окружающий его контекст в виде отношений, в которые вступает проверяемое понятие, является частным случаем контекста, в котором применяется базовое понятие в онтологии.

3. $\exists o=(X,t)\in M_o : X\neq X_i \wedge t_i \subseteq t \wedge O_t(o_i)\subseteq O_o(o)$, аналогично предыдущему варианту с той лишь разницей, что окружающий контекст представлен смежными понятиями, с которыми связано проверяемое и базовое понятие некоторыми отношениями (суть самих отношений не проверяется).

4. $\exists o=(X,t)\in M_o : X\neq X_i \wedge t_i \subseteq t \wedge O_t(o_i)\subseteq O_o(o) \wedge L_t(o_i)\subseteq L_o(o)$, факт наличия противоречия устанавливается в случае, когда окружающий проверяемое понятие контекст представлен как отношениями, так и понятиями с которыми связано o_i. Противоречие обнаруживается только тогда, когда все элементы контекста понятия o_i будут обнаружены в контексте базового понятия онтологии.

4.1.3. Использование неопределенного понятия

Имеется семантическое представление M_t анализируемого текста вида (5). Понятие $o_i=(X_i,t_i)\in M_t$ считается неопределенным в онтологии M_o, если истинно одного из следующих высказываний.

1. $\neg\exists o=(X,t)\in M_o : X=X_i \wedge t_i \subseteq t$, в проверяемом тексте используется понятие o_i, не заданное в онтологии.

2. $\exists o=(X,t)\in M_o : X=X_i \wedge t\subseteq t_i$, в онтологии существует понятие o в таком виде, что смысловое содержание проверяемого понятия o_i полнее имеющегося в онтологии смыслового содержания понятия o (набор значений у анализируемого понятия больше, чем набор значений понятия в онтологии). Данный вариант позволяет сделать вывод о том что, либо онтология неполноценна, либо анализируемый документ относится к другой предметной области, в которой o является действительным базовым понятием, а в данной же предметной области o может выступать только в роли контекста действительным базовым понятиям.

4.2. Противоречия в предикатах

Имеется семантическое представление M_t анализируемого текста вида (5). Некоторое отношение $l_i \in L_t$ связывает понятия O_i данного представления. Истинность одного из следующих высказываний указывает на неправильное использование отношения.

1. $\forall o\in O_i \ \neg\exists l\in L_o(o) : l=l_i$, множество всех отношений, в которые могут вступать понятия, связанные посредством l_i в анализируемом тексте, не содержит самого l_i в онтологии.

2. $L_t(O_i)\not\subset L_o(O_i) \wedge O_t(O_i)\not\subset O_o(O_i)$, менее строгое ограничение, выявляющая несоответствие только в том случае, если онтологией не предусмотрена вся группа понятий O_i, связанная отношением l_i.

Существует так же нарушение порядка применения предикатов. Оно подразумевает наличие условий применения предикатов, которые отражены контекстным окружением (инцидентные понятия и смежные ребра), заданным онтологией для проверяемых предикатов. В случае нарушения в анализируемом тексте будет обнаружено контекстное окружение некоторого отношения не включенное в окружение этого же отношения в онтологии.

5. Поиск противоречий в ЭБ «Мониторинг правового пространства и правоприменительной практики в Совете Федерации»

Рассмотренные модели и методы можно использовать для создания ЭБ в любой предметной области, для которой может быть сформирован базовый корпус текстов, закладываемый в ее модель и онтологию. К такому корпусу предъявляется ряд требований таких как, непротиворечивость, полнота, конечность. Будем считать, что в действующем законодательстве можно выделить корпус нормативно правовых актов, удовлетворяющий этим требованиям. Поэтому логичным является построение системы поиска противоречий (СПП), облегчающей как законотворческую деятельность, так и анализ результатов законотворчества. Последний представляется возможным реализовать и внедрить уже в настоящее время в рамках разрабатываемой ЭБ «Мониторинг правового пространства и правоприменительной практики в Совете Федерации». Далее коротко рассмотрены основные аспекты этой системы,

а так же изложены варианты использования СПП в рамках одной из ее подсистем.

5.1. Назначение, функции и структура ИС

Система «Мониторинг» необходима для автоматизации сбора, анализа и обобщения информации о качестве реализации конституционных полномочий Совета Федерации (подробнее см. [10]). Кроме традиционных задач сбора, накопления и поиска информации перед данной ИС поставлены следующие:
1. анализ накопленных данных (например, на предмет соответствия законодательству), обобщение и систематизация информации;
2. проведение правовой, лингвистической и научной экспертизы, ведение мониторинга правоприменительной практики;
3. создание результирующих документов аналитического характера (синтез новых знаний).

Обобщенная структура системы (см. рис. 8) включает подсистему «Анализ», которая и предназначена для решения указанных задач.

5.2. Структура ИС «Мониторинг»

На рис. 8 приведена структура ИС «Мониторинг» с указанием направлений потоков передаваемой информации.

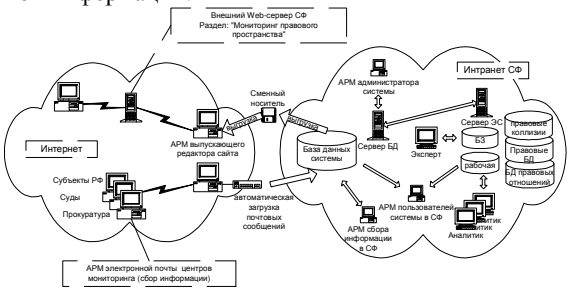

Рис. 8. Структура ИС «Мониторинг»

Важным добавленным компонентом является подсистема «Анализ», включающая в себя сервер экспертной системы (ЭС), рабочее место эксперта и места аналитиков. Основное назначение сервера ЭС – поддержка базы знаний (БЗ) экспертной системы, в которой заложены модели и онтологии предметных областей, соответствующих различным правовым сферам. Корректировку БЗ выполняет эксперт. Аналитики являются пользователями подсистемы с точки зрения функций, приведенных в предыдущем параграфе.

5.3. Подсистема «Анализ»

Выявление противоречий, а так же такая обработка текстового материала как систематизация и обобщение, необходимо в рамках всех задач приведенного ранее перечня. Поэтому в подсистеме «Анализ» целесообразно использовать модули семантического анализа и поиска противоречий.

Рис. 9 иллюстрирует вариант использования модулей для выявления противоречия документа действующему законодательству. Рабочий материал подвергается семантическому анализу, после чего полученное семантическое представление текста вида (5) сопоставляется с онтологией, заложенной в базу знаний (БЗ), при этом используются формальные правила, изложенные ранее. В случае выявления противоречивых фрагментов пользователь получает ссылки на соответствующий участок документа и законодательства РФ, представляющей корпус текстов БЗ.

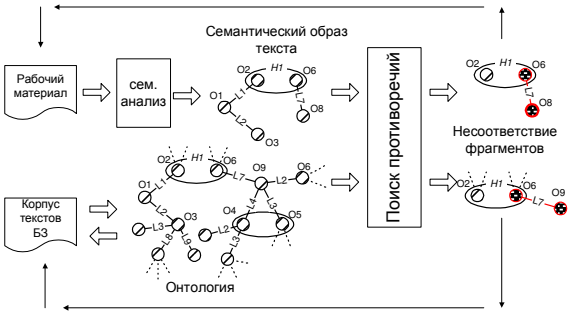

Рис. 9. Поиск противоречий в рабочих документах.

На рис. 10 приведена иллюстрация работы модуля для обнаружения противоречия во фрагменте текста: «Субъект права обязан вносить плату за регистрацию прав на недвижимое имущество». В данном случае используется фрагмент онтологии, получение которого рассмотрено ранее (см. рис. 6). В качестве правила, выявляющего это противоречие, может быть использовано либо правило выявления неопределенного понятия, либо правило нарушения предиката. Первое сработает, если в онтологии вообще не определено понятие «плата за регистрацию прав не недвижимое имущество», второе будет использоваться в том случае, если указанное понятие присутствует в онтологии, но использование отношения ПЛАТИТЬ(…) к нему не применимо. В обоих случаях модуль укажет на понятие, в котором возникло несоответствие (см. рис. 10).

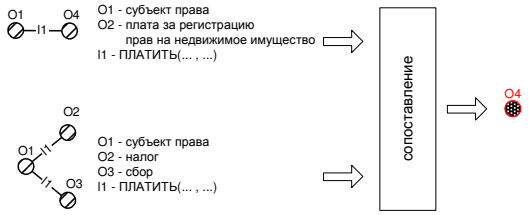

Рис. 10. Сопоставление с онтологией.

Кроме несоответствий, выявляемых согласно описанным формальным правилам, возможен вариант неточного высказывания. В качестве примера неточного высказывания рассмотрим фрагмент текста: “Средства, получаемые в виде платы за регистрацию и предоставление указанной информации, используются исключительно на создание, поддержание и развитие системы государственной регистрации прав на недвижимое имущество.” На рис. 11 приведена семантическая сеть, построенная для данного предложения.

Неточность высказывания заключается в том, что в предложении не указан субъект, выполняющий платеж, а так же не указан явный получатель платежа. С точки зрения русского языка никаких ошибок не присутствует, но с точки зрения пред-

метной области, а именно законодательства, предложение неточно.

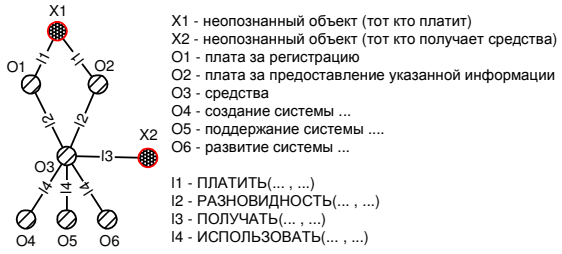

Рис. 11. Не доопределенная семантическая сеть.

Семантический анализатор достроит необходимые узлы-понятия (X1, X2 на рис. 11), но эти понятия будут «не достроены», поскольку слов, указывающих на них в тексте не обнаружено. Именно этот факт и позволит заострить внимание пользователя на данном фрагменте текста.

5.4. Обобщенное функционирование модуля поиска противоречий

Обобщенно модуля поиска противоречий функционирует следующим образом. Модуль перебирает все формальные правила выявления противоречий, подобные рассмотренным ранее. В соответствии со спецификой каждого правила из входного семантического представления текста извлекаются фрагменты: понятия, отношения и их связки. Далее выполняется поиск извлеченных фрагментов в онтологии. Образуются пары из найденных элементов онтологии и фрагментов входной структуры и выполняется их сопоставление в соответствии с текущим правилом. На выходе модуля выдаются пары, не удовлетворяющие текущему правилу.

Заключение

В работе разработана структура модели и онтологии предметной области, применение которой продемонстрировано на задаче поиска противоречий в правовых документах. В основе разработки лежат более ранние подходы к построению системы понимания текстов [2]. Важным аспектом подобных систем является эффективные методы наполнения модели и онтологии с разработанной здесь структурой. В статье показана возможность автоматического наполнения онтологии, но вопрос автоматизации наполнения модели остается открытым. Ему планируется посвятить ряд последующих работ.

Возможности системы понимания текстов продемонстрированы на примере внедрения в развивающийся проект ЭБ «Мониторинг правового пространства и правоприменительной практики в Совете Федерации», выполняемая по заказу СФ РФ.

Литература

[1] Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К. В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа. 5-ая Всероссийская научная конференция RCDL'2003.

[2] Андреев А.М., Березкин Д.В., Симаков К.В. Архитектура системы машинного понимания текстов. Сборник трудов 1. Информатика и системы управления в XXI веке. М.: Изд-во МГТУ им. Н.Э. Бауман, 2003.

[3] Андреев А.М., Березкин Д.В., Симаков К. В. Формальный V - язык описания морфологии и синтаксиса текстов на естественном языке. Сборник трудов 1. Информатика и системы управления в XXI веке. М.: Изд-во МГТУ им. Н.Э. Бауман, 2003.

[4] Андреев А.М., Березкин Д.В., Симаков К.В. Снятие синтаксической омонимии в задачах машинного понимания естественных текстов. Сборник трудов 1. Информатика и системы управления в XXI веке. М.: Изд-во МГТУ им. Н.Э. Бауман, 2003.

[5] Козеренко Е.Б. Концептуально-лингвистическое моделирование в интеллектуальных системах на основе расширенных семантических сетей: Автореферат диссертации на соискание ученой степени к. фил. наук: – М., 1995.

[6] Кузина Л.Н. Автоматизированное формирование семантических моделей сложных объектов по текстовым источникам: Автореферат диссертации на соискание ученой степени к. физмат наук: – М., 1996.

[7] Леонтьева Н.Н. К теории автоматического понимания естественных текстов. Ч.2: Семантические словари: состав, структура, методика создания – М.: Изд-во МГУ, 2001

[8] Нариньяни А.С., Кентавр по имени ТЕОН: Тезаурус + Онтология, Российский НИИ искусственного интеллекта. (http://www.artint.ru/articles/narin/teon.htm)

[9] Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. – М.: Наука, 1989.

[10] Проблемы информационно-технологического сопровождения процесса мониторинга правового пространства и правоприменительной практики в Совете Федерации Федерального Собрания - М.: Совет Федерации, 2004.

[11] Тейз А., Грибомон П., Юлен Г. Логический подход к искусственному интеллекту. От модальной логики к логике баз данных. – М.: Мир, 1998.

[12] Тузов Математическая модель языка. – СПб.: Изд-во СПбГУ, 1984.

[13] Уэно Х., Кояма Т., Окамото Т. и др. Представление и использование знаний. – М.: Мир, 1989.

[14] Automatic acquisition of phrasal knowledge for English-Chinese bilingual information retrieval. Ming-Jer Lee, Lee-Feng Chein ACM SIGIR 1998.

[15] Automated message understanding: a real world prototype. Jenkins T., Gaillard A., Holmback H. etc. ACM 1990.

[16] Domain specific knowledge acquisition from text. Moldovan D., Girju R., Rus V. ACM 2002.

[17] Knowledge acquisition from prescriptive texts. Moulin B., Rousseau D. ACM 1990.

[18] Ontological Modeling. L. Kalinichenko, M. Missikoff, F. Schiappelli. 5-ая Всероссийская научная конференция RCDL'2003.

[19] Ontological Semantics, Formal Ontology, and Ambiguity Nrenburg S. Raskin V. ACM 2001.

[20] PALKA: A system for lexical knowledge acquisition. Kim J.T., Moldovan D. I. ACM SIGART 1993.

[21] The automatic initialization of an object-oriented knowledge base. Cordova J.L., Hodges J.E. ACM 1992.

The peculiarities of domain's model and ontology development for ambiguity detection in legal digital libraries

This paper is devoted to a problem of ambiguity and contradictions detection in natural language texts from legal purview. In our approach knowledge base consists of two major parts: domain's model and domain's ontology. The first one represents general properties of domain, while the second one collects base concepts and relationships and is used in such specific applications as ambiguity detection. We show how ontology could be obtained by automatic processing of representative corpus. We present the structure of model and ontology from our point of view and appropriate methods of its treating. The formal models of contradictions are also presented.