УДК 681.3

О.В. Лесько, Ю.В. Рогушина

АНАЛИЗ СЕМАНТИКИ ЕСТЕСТВНЕННО-ЯЗЫКОВЫХ ЗАКОНОДАТЕЛЬНЫХ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

В статье анализируются проблемы, возникающие в процессе поиска информации, интересующей пользователя, в законодательных документах, и обосновывается необходимость интеллектуализации такого поиска при помощи использования лингвистических ресурсов и онтологических баз знаний соответствующих предметных областей. Предлагаются структура и метод построения онтологии предметной области и связанной с ней системы лингвистического анализа, обеспечивающие как выполнение семантической разметки законодательных документов персонализированной терминологией пользователя, так и выполнение семантического поиска, предусматривающего выполнение расширенных и модифицированных запросов к ним. При этом используется связь между терминами онтологии и фрагментами естественно-языкового текста и основанная на ней нормализация терминологии предметной области. Описывается программная реализация информационной системы, основанной на описанных в работе моделях и методах и обеспечивающей семантический поиск в законодательных документах.
Ключевые слова: онтология, поиск информации, семантическая разметка.

Введение

Увеличение объема, динамичность и усложнение структуры естественно-языкового контента Web требует развития автоматизированных средств и методов извлечения знаний из текстов, относящихся к различным предметным областям (ПрО). Наиболее эффективны такие подходы в тех сферах деятельности, где используются достаточно специфичный (со значениями, часто отличающимися от общеупотребимой лексики), относительно устоявшийся и формализованный набор терминов и достаточно сложные отношения между этими терминами, к которым целесообразно применять правила логического вывода. Одной из таких предметных областей является законодательство – в таких документах излагаются формально непротиворечивые правила, которые пользователю надо интерпретировать к своим данным.

Актуальность проблемы

В настоящее время значительная часть законодательных документов доступна через Web. Тем не менее, пользователям, не являющимся юристами: крайне сложно найти в этих документах ответ на интересующий их вопрос, так как они, как правило, не владеют соответствующей терминологией, не имеют сведений о структуре данных документов и поэтому не могут корректно сформулировать запрос.

Это вызывает необходимость в разработке интеллектуальных средств, обеспечивающих преобразование запроса на естественном языке в произвольной форме в формальный запрос, состоящий из терминов ПрО и учитывающий семантические связи между этими терминами (в особенности – родовидовые отношения и отношение синонимии).

При этом опыт разработки современных интеллектуальных Web-приложений свидетельствует о том, что знания о предметной области целесообразно представить в виде онтологии – это обеспечивает их интероперабельность и наличие инструментальных средств для редактирования и визуализации.

Кроме того, возникает необходимость в использовании лингвистических знаний для сопоставления естественно-языковых (ЕЯ) фрагментов текста с формальными понятиями, представленными в виде экземпляров и классов онтологии ПрО.

Использование лингвистических ресурсов в системах семантического поиска

Современные информационно-поисковые и информационно-аналитические системы работают с текстовой информацией в широких или неограниченных предметных областях, т. е. областях, в состав которых входят тысячи разных классов сущностей, входящих между собой в неограниченные типы отношений. Поэтому характерной чертой современных методов обработки текстовой информации в таких системах стало минимальное использование знаний о мире и о языке, опора на статистические методы учёта частотностей встречаемости слов в предложении, тексте, наборе документов, совместной встречаемости слов и т. п. Учитывая это, когда подобные операции выполняет человек, ему необходимо выявить основное содержание документа, его основную тему и подтемы, и для этого обычно используется большой объём знаний о языке, о мире и об организации связного текста.

Внедрение в современные интеллектуальные системы методов автоматической обработки текстовой информации является сложной задачей. Это обусловлено тем, что используемые при этом лингвистические знания должны описываться в специально создаваемых ресурсах и содержать описания десятков тысяч слов и словосочетаний. Для эффективного применения таких ресурсов нужно, кроме того, обеспечить поддержку логического вывод и разрешение многозначности естественного языка.

Сегодня наиболее широко используются три подхода к представлению лингвистических знаний:

- традиционные информационно-поисковые тезаурусы;
- тезаурус WordNet;
- специализированные формальные онтологии.

Исторически первыми для таких задач применялись традиционные информационно-поисковые тезаурусы, которые создавались для ручного индексирования документов на основе национальных и международных стандартов и были не совсем пригодны для автоматизированной обработки информации в электронной форме.

Тезаурус WordNet был разработан в 90-е годы 20-го века в Принстонском университете США и представлял собой иерархическую сеть лексических понятий английского языка. WordNet 3.0 содержит 155 тысяч лексем и словосочетаний, которые организованы в 117 тысяч понятий. Разработаны его версии для многих языков. Если информационно-поисковые тезаурусы описывают определенную ПрО, то WordNet содержит сведения об общей лексике того или иного языка (хотя можно строить тезаурусы типа WordNet и для конкретных ПрО). Следует отметить, что структура WordNet не приспособлена для описания терминологии ПрО: раздельное описание частей речи, большое количество не связанных между собой лексем, слабая поддержка обработки словосочетаний вызывают многочисленные проблемы при практическом использовании таких тезаурусов.

Базовым понятием в WordNet является лексема, а базовым отношением – отношение синонимии. В состав словаря входят лексемы, каждая из которых относится к одной из четырех категорий (соответствующих частям речи): существительное, прилагательное, глагол и наречие. С каждой из частей речи связан свой набор отношений. Например, между существительными могут существовать отношения синонимии, антонимии, меронимии и гипонимии.

Наборы синонимов (синсеты) – это основные структурные единицы WordNet. Два выражения считаются в WordNet синонимами, если замена одного из них на другое в высказывании не меняет значения истинности этого высказывания.

Основное отношение между синсетами – родо-видовое, при этом видовой синсет называется гипонимом (А является гипоимом Б, если истинно утверждение “А является разновидностью Б”), а родо-

вой – гиперонимом. Отношения между синсетами образовывают иерархическую структуру.

В настоящее время для представления баз знаний самых разных ПрО все чаще используются онтологии. Для целей автоматизированной обработки ЕЯ-текстов разрабатываются специализированные онтологии, как правило, не полностью определяющиеся в терминах формальных свойств и аксиом, т. е. легкие онтологии. Эти онтологии объединяют принципы разработки традиционных информационно-поисковых тезаурусов и лингвистических ресурсов типа WordNet с методологиями создания формальных онтологий.

В частности, в [1] приводится формальная модель лингвистической онтологии для широкой ПрО

$$O_{lingv} =$$

$$= \langle C, Ex, NO, R_{lingv}, A_{tr,i}, S, T, M_{m,a}, L, DC \rangle,$$

где

C – множество понятий онтологии O_{lingv}, где понятие обозначает класс сущностей, обладающих одинаковыми свойствами и отношениями с другими классами сущностей;

Ex – множество экземпляров понятий онтологии O_{lingv}, такое, что задано отображение $E: C \to 2^{Ex}$;

NO – множество уникальных имен понятий и экземпляров в онтологии O_{lingv};

R_{lingv} – набор отношений между понятиями $R_{lingv} \subseteq C \times C$, который специально сформирован для автоматической обработки текстов;

$A_{tr,i}$ – множество правил вывода, которые базируются на свойствах транзитивности и наследования отношений;

S – множество отношений между языковыми выражениями T и понятиями C: $\{s(c_i, t_j)\}$;

T – множество текстовых входов онтологии – языковых выражений, значения которых представлены в онтологии O_{lingv};

$M_{m,a}$ – множество многозначных слов и выражений из T: $M_{m,a} \subseteq T$, при этом многозначные входы онтологии делятся на два подвида: M_m – текстовые входы, которые относятся к более чем одному понятию онтологии, M_a – текстовые входы, которые многозначны, но в онтологии O_{lingv} для них представлено только одно значение, $M_{m,a} = M_m \cup M_a$;

L – множество лемматических представлений языкового выражения, т. е. представление выражения в виде последовательности слов в словарной форме (например, словосочетание “сельское хозяйство” представляется в лемматическом виде как “сельский хозяйство”;

DC – отображение терминологического состава TD заданной коллекции ПрО $Dcoll$ на текстовые входы и понятия онтологии $DC: (Dcoll, TD) \to (T, C)$, которое задает критерий минимальной полноты онтологии, который должен обеспечивать покрытие терминологического состава заданной коллекции предметной области.

В текстах ПрО значительную часть составляют слова, которые не являются специфичными для этой конкретной ПрО, т. е. принадлежат общему лексикону GL. Поэтому многозначные слова делятся на два множества. Множество M_m содержит слова, которые могут быть отнесены к более чем двум понятиям O_{lingv}, а в множество M_a входят те слова, которые связаны с различными значениями в GL.

Таким образом, *лингвистическая онтология* ПрО представляет собой БЗ онтологического типа о понятийной системе и лексико-терминологическом составе ПрО.

В работе [2] рассматривается модель лингвистической онтологии для

автоматической обработки текстов предметной области, в состав которой входят тысячи разных классов сущностей, имеющих между собой неограниченные типы отношений и ситуаций. По мнению авторов, предложенная система отношений отражает наиболее существенные взаимосвязи между сущностями и может применяться для описания отношений между понятиями в самых разных предметных областях.

Анализ смысла естественно-языковых текстов со сложной структурой относительно ограниченной предметной области.

Под анализом смысла будем понимать проверку истинности утверждений, связывающих несколько терминов ПрО отношениями из ограниченного подмножества. При этом генерация самих утверждений является прерогативой пользователя. Для того, чтобы предоставить пользователю терминологическую базу для запросов, на основе лингвистического анализа ЕЯ-текстов ПрО формируется соответствующая онтология.

В дальнейшем пользователь может использовать запросы на ЕЯ, которые будут интерпретированы анализатором словоформ (oly-анализатор).

В онтологии пользователь может отслеживать иерархию терминов и их синонимию, а семантическая разметка позволяет обнаруживать нужные фрагменты текста.

Это связано с тем, что у многих пользователей в процессе поиска возникают проблемы, связанные с незнанием терминологии ПрО, что не позволяет создавать корректные запросы по интересующим пользователя запросам.

В данной работе представляется целесообразным разделить знания о ПрО на две части – онтологию ПрО, которая отображает основные понятия и связи этой области, и лексическую онтологию, включающую лингвистические сведения о тех словах и словосочетаниях, которые используются в ЕЯ-документах, релевантных ПрО.

Семантический поиск

Обнаружение знаний в Web является составной частью многих интеллектуальных приложений.

В работе [3] предложено *моделирование системы взаимодействия* между *ИР* и *потребителями* информации, которая привлекает к этому процессу внешние и внутренние *базы знаний* и обеспечивает логический вывод на этих знаниях в открытой гетерогенной информационной среде Web, рассмотрено использование этой модели при решении прикладных задач.

В наиболее обобщенном понимании *информационный поиск* – сложная проблема *сопоставления* представления пользователя о нужных ему знаний с контентом доступных ИР и *построения* на основе этого сопоставления информационного объекта (ИО) с конечным набором свойств, значения которых извлекаются из этих ИР.

Семантический поиск – это информационный поиск, в котором такое сопоставление и построение ИО выполняются на семантическом уровне, т. е. с использованием знаний.

Пользователь имеет часть информации об ИР и пытается дополнить ее сведениями, извлеченных из различных источников.

Основное отличие семантического поиска от традиционного – использование *знаний* об объекте поиска, пользователях, ИР и предметной области (ПрО) поиска.

Семантический поиск – *комплексная* научная задача, основанная на таких достижениях в области искусственного интеллекта, как общая теория представления и обработки знаний, распознавание образов, логический вывод.

Метод построения онтологии предметной области на основе ЕЯ-документа (на примере Налогового кодекса) и толкового словаря

Структура онтологии ПрО

Онтология ПрО разрабатывается таким образом, чтобы обеспечить поддержку семантического поиска в корпусе

законодательных документов – например, в Налоговом кодексе.

Поэтому в ней присутствуют:

– термины, специфичные для данной ПрО;

– некоторое количество общеупотребимых терминов, необходимых для однозначного определения контекста использования терминов ПрО (например, “месяц”, “год”, “сумма”, “процент”), которые могут быть определены либо непосредственно, либо путем ссылки на внешние онтологии – как верхнего уровня, так и специализированные;

– иерархические отношения между терминами онтологии ПрО – различные типы мериологических отношений “часть-целое”, отношение “класс-подкласс”, отношение “класс-экземпляр”;

– отношения синонимии, позволяющие расширить терминологический словарь, используемый для поиска;

– отношения, специфичные для данной ПрО, для описания семантики которых используется анализ соответствующих статей толкового словаря.

Таким образом, для представления ПрО используется “легковесная” онтология, не содержащая аксиом. Это значительно упрощает ее использование и обеспечивает более быструю работу алгоритмов ее анализа.

Такая онтология обеспечит поиск не только по заданному ключевому слову, но и по словам, связанным с ним какими-либо отношениями (в данной версии используются только таксономические отношения «класс-подкласс»). Так, если задано слово "транспорт", то поиск должен происходить и для терминов "автомобиль", "самолет", "грузовик", "самокат".

На первом этапе строится терминологический словарь ПрО. Для этого из релевантного ЕЯ-текста извлекаются все именные группы (существительные). Из этого списка эксперт по знаниям вручную выбирает те именные группы, которые специфичны для ПрО: с учетом специфики составления законодательных документов, специфичны (т. е. необходимы для корректного описания семантики документа) практически все встречающиеся в нем именные группы. Именная группа представляет собой несколько стоящих рядом существительных и прилагательных. Основные методы выделения и анализа именных групп описаны в [4]. При этом основная задача эксперта заключается в том, чтобы отделить имена классов от имен экземпляров (при этом практически все поименованные сущности должны быть отнесены к экземплярам). Процедуры поиска названий классов и поименованных сущностей приведены в [5]. В общем виде это сводится к выполнению следующих шагов.

Если для слова wi и соответствующего класса ci, $\exists$ o, o $\in$ ci и o $\in$ ci-1, то выделяется соответствующая именная группа.

Следует отметить, что в законодательные документы (в частности, в Налоговый кодекс) постоянно вносятся поправки и дополнения. Поэтому при каждом изменении обрабатываемого документа необходимо проверять, появились ли в нем новые термины. Если такие термины обнаружены, то их надо включить в онтологию, а для описания их семантики проанализировать соответствующую статью толкового словаря. Таким образом, первый этап повторяется итеративно при всех изменениях исходного обрабатываемого документа, а онтология ПрО модифицируется в соответствии с обнаруженными изменениями.

На втором этапе формируется список отношений ПрО – аналогично списку терминов, но из текста извлекаются глаголы. Для каждого отношения уточняется лишь, к которой из трех категорий оно относится – иерархические, синонимии, специфичные для ПрО. Основные типы отношений, их характеристики и способы описания приведены в [6].

Целью такого упрощения семантики является увеличение скорости обработки ЕЯ-текстов. Как показывает практика, такое упрощение почти не снижает качество поиска.

На третьем этапе для каждого из терминов в толковом словаре находится определение, которое описывает семантику этого термина и его связи с другими терминами онтологии ПрО. Анализ таких определений позволяет дополнительно ввести в онтологию отношения между терминами – в первую очередь, родовидовые и синонимические, а также атрибуты (свойства) этого класса и их возможные значения.

Таким образом, впоследствии при поиске или анализе текста можно обнаруживать не только заданный термин, но и его подклассы, экземпляры, надкласс или термины-синонимы.

Извлечение терминов из ЕЯ-документов. Так как в толковом словаре каждая статья представляет собой фрагмент ЕЯ-текста, а сам анализируемый законодательный документ – ЕЯ-текст. Возникает необходимость в средствах лингвистического анализа, позволяющих использовать знания о конкретном естественном языке для выделения отдельных лексем и словоформ, связанных с терминами и отношениями онтологии ПрО.

Именно для этого предназначена лексическая онтология, более детально описанная в [7].

Извлечение семантики терминов онтологии на основе определений из толкового словаря

Для выделения специфичных для ПрО "Налоговый кодекс" терминов и отношений между ними применяется следующий алгоритм.

Вначале в тексте Налогового кодекса выделяются все существительные, производные от существительных прилагательные и глаголы.

Затем каждое из найденных слов преобразуется в исходную форму (например, для существительного это именительный падеж единственного числа, для глагола – инфинитив и т. д.).

После выполнения такой нормализации для каждого найденного слова в толковом словаре (предполагается, что толковый словарь, релевантный ПрО, доступен системе анализа, и его статьи могут обрабатываться средствами этой системы) надо найти его определение – словарную статью, заглавие которой совпадает с анализируемым термином ПрО, – и выполнить анализ найденного определения следующим образом:

1) выделяется первая группа существительного, идущая после описываемого понятия;

- для каждого слова группы слова "група", "клас", "вид", "тип", «галузь», "сукупність" заменяются именем онтологического иерархического отношения «подкласс», связывающий это нормализованное существительное с определяемым термином («свинарство – галузь тваринництва» трансформируется в «свинарство” «подкласс» «тваринництво»);

- если слово стоит в именительном падеже, то оно считается вышестоящим классом («свиня – тварина» трансформируется в «свиня» «подкласс» «тварина»);

- если слово стоит в родительном падеже, то оно связывается с исходным словом онтологическим отношением «связано» («Аналіз – це пошук сутностей» трансформируется в «пошук» «связано» «сутність»);

2) выделяется глагол; в онтологию добавляется соответствующее отношение между терминами, соответствующими существительным перед и после глагола (“корова дає молоко” трансформируется в “корова” «связано» «молоко»;

3) выделяется следующая группа существительных, в ней выделяется главное слово (любой падеж, кроме винительного, "вирощувати вовну білих овець").

Обработка продолжается до тех пор, пока не обработаны все части словарной статьи.

Аналогично выполняется обработка для всех существительных из налогового кодекса, а затем полученные триплеты «субъект-отношение-объект» объединяются в одну онтологию ПрО. Иерархические отношения позволяют организовать таксономию терминов ПрО.

При этом пользователь может включать в онтологию только те термины и отношения, которые он считает существенными.

Использование онтологии ПрО для семантической разметки документов ПрО

Семантическая разметка ЕЯ-текста заключается в том, чтобы установить связи элементов текста – слов, словосочетаний, предложений – с некоторыми понятиями соответствующей ПрО. Связи в общем случае устанавливаются по принципу "многие со многими", а сама разметка зависит от выбранной ПрО и способа формализации знаний соответствующей ПрО.

Семантическая разметка текстов позволяет автоматизировано анализировать их в дальнейшем на смысловом уровне, выполнять над текстами различные логические операции, извлекать из них новые знания и т. п. При формировании семантической разметки нужно использовать не только знания ПрО (или хотя бы ее терминологическую базу), но и правила того конкретного естественного языка, на котором написан текст. К сожалению, создание такой разметки является нетривиальной и довольно трудоемкой задачей. Семантическая разметка зависит и от того, какие именно средства используются для описания ПрО.

Формально семантическая разметка произвольного текста может быть определена следующим образом:

Текст X, $X = \langle x_1,...,x_n \rangle$ представляет собой конечную последовательность символов, принадлежащих конечному множеству A, $\forall, i = \overline{1, n}, \quad x_i \in A$.

При этом часть символов являются разделителями (символами, отделяющими дуг от друга отдельные слова текста) и относятся к множеству B, $B \subseteq A$. Примеры разделителей пробел, точка, запятая.

ПрО, для которой осуществляется семантическая разметка, характеризуется набором терминов из конечного множества T, $T = \{t_1,...,t_m\}$. Эти термины могут использоваться в качестве тэгов семантической разметки.

Произвольный фрагмент текста

$$\langle t_p,...,t_q \rangle, \; p,q = \overline{1, n}, \quad p < q$$

может быть связан с одним или несколькими понятиями из Т. Для этого соответствующим тэгом отмечают начало такого фрагмента, а парным ему закрывающим тэгом – конец фрагмента.

Семантическая разметка текстов позволяет автоматизировано анализировать их в дальнейшем на семантическом уровне, выполнять над ними разные логические операции, извлекать из них новые знания и т. п.

При формировании семантической разметки нужно использовать не только знания ПрО (или хотя бы ее терминологическую базу), но и правила того конкретного естественного языка, на котором написан текст. К сожалению, создание такой разметки является нетривиальной и довольно трудоемкой задачей. Семантическая разметка зависит и от того, какие именно средства используются для описания ПрО.

Алгоритм семантической разметки текстов. Семантическая разметка ЕЯ-текстов для определенной ПрО создается в два этапа. На первом этапе производится обучение с помощью алгоритма накопления лингвистических сведений о ПрО (АНЛС). Необходимо сформировать следующие множества:

1) P_w – словоформы, связанные с понятиями онтологии ПрО. Эта информация может быть извлечена из различных словарей синонимов, лингвистических баз данных, а также явным образом вручную из корпуса текстов;

2) R_w – словоформы, связанные с отношениями онтологии ПрО (аналогично);

3) I – отношения именования (ОИ), связывающие: а – ПС и классы, б – классы и подклассы;

4) I_w – словоформы, связанные с ОИ;

5) шаблоны, связывающие ПС и имена их классов (в общем случае слабо зависящие или вообще не зависящие от предметной области). Эта операция может быть выполнена один раз, но в дальнейшем множество шаблонов может расширяться для учета специфики ПрО. Каждый шаблон представляет собой строку символов, состоящую из имени предиката и модели управления, например, «называется».

В корпусе текстов находятся слова, написанные с большой буквы и не входящие в общий словарь, состоящие из больших букв и слова, взятые в кавычки, и для них выделяются синтаксические шаблоны, определяющие указание на принадлежность ПС к определенному классу. Затем в предложении с такими ПС обнаруживаются имена классов, к которым принадлежат эти ПС. Если такое имя класса присутствует, то осуществляется попытка выделить слова, связывающие синтаксически ПС и имя ее класса. Если это удается, то для этих слов – ОИ– строится шаблон.

На вход АНЛС подаются: 1) онтология О, характеризующая знания пользователя об интересующей его ПрО; 2) словарь лексем естественного языка; 3) обучающая выборка ЕЯ-текстов, по которой при участи пользователя формируется набор словоформ, связанных с терминами онтологии О (корпус текстов).

После того, как формируется набор множеств слов ЕЯ-текстов, каждый из которых соответствует определенному термину онтологии

$$\forall\ t_i \in T,\ \ t=\overline{1,n},\ \ \exists\ S_i=\{s_{i_1},...,s_{i_m}\},$$

эти множества могут быть преобразованы для более эффективной обработки текста. Часть элементов множества

$$S_i=\{s_{i_1},...,s_{i_m}\}$$

могут быть опознаны пользователем как одна словоформа и заменены элементом, являющимся общей частью этих элементов,

$$\exists\ l, l\subseteq s_{i_k},\ k=\overline{1,p},\ p\geq 2,\ s_{,i_k}\in S_i.$$

В результате обучения системы каждому термину онтологии О приписывается 0, 1 или несколько словоформ, соответствующих в ЕЯ данному понятию. Словоформы извлекаются из обучающего множества ЕЯ-текстов, отнесенных пользователем к определенной ПрО, описанной в О с точки зрения информационных интересов пользователя. Полученная информация заносится в таблицу С.

Семантическую разметку законодательного документа при необходимости можно выполнять и в процессе формирования онтологии ПрО (до ее окончательного построения) с целью выделения важных закономерностей ПрО.

Разработка лексической онтологии ПрО для поддержки поиска фрагментов ЕЯ-текста, соответствующих терминам онтологии ПрО

Лексическая онтология – это простой тезаурус, который содержит термины ПрО. Также в лексической онтологии содержатся ссылки на соответствующие классы онтологии предметной области пользователя. Словоформы могут быть построены автоматически на основе какой-либо лингвистической базы данных (типа орфографического словаря) либо заполнены вручную пользователем – для новых или специфичных терминов ПрО. O_L пополняется итеративно: при каждом добавлении к онтологии O_1 нового класса в O_L также включается этот новый класс, для которого тем или иным образом формируется словоформа.

O_1 пополняется классом t_i, если в одном абзаце текста из T_0 есть связанные с ним фрагменты, а также словоформы другого класса, входящего в O_1 и отношения.

Новые классы для O_1 извлекаются из информации, хранящейся в текстах T_0

путем их лингвистического анализа. Например, для класса «собака» могут быть найдены такие супер классы, как «млекопитающие», «животные», подклассы как «стаффордширский терьер», синонимы – с точки зрения конкретного пользователя (например, «живая подушка»), атрибуты («лапы», «хвостик», «цвет»). При этом сам пользователь может принять решение о том, что считать подклассами, а что – атрибутами.

Для этого в лексическую онтологию помещаются имена отношений ПрО и соответствующие им словоформы. Например, отношению «состоит из» соответствуют такие фрагменты ЕЯ-текста, как «сделан из», «изготовлен из», «включает в себя». Такие словоформы могут быть сформированы следующим образом: если в одном предложении ЕЯ-текста обнаружены два фрагмента $t_1 \in O_L$ и $t_2 \in O_L$, которые являются экземплярами классов лексической онтологии O_L s_1 и s_2 соответственно, которым соответствуют классы онтологии ПрО O_1 c_1 и c_2 такими, что между этими классами O_1 есть отношение r_i, то тот фрагмент ЕЯ-текста q_i, который находится между фрагментами t_1 и t_2, может являться экземпляром словоформы r_i.

И наоборот, если в одном предложении встретились два фрагмента $t_1 \in O_L$ и $t_2 \in O_L$, которые являются экземплярами классов лексической онтологии O_L s_1 и s_2 соответственно, но в онтологии ПрО не зафиксированы отношения между этими классами, то необходимо спросить пользователя о необходимости пополнения онтологии ПрО новым отношением.

Использование лексической онтологии позволяет избежать хранения в онтологии длинных названий типа «столы компьютерные угловые с надставкой типа шкаф и двумя полками», которые сложно обрабатывать, поскольку может измениться порядок слов.

Использование внешних лингвистических баз данных для генерации лингвистической базы данных украинского языка

При создании лексической онтологии были использованы лингвистические знания о парадигмах словоформ украинского языка, представленные на интернет-сайте “Українського мовно-інформаційного фонду НАН України” (http://www.ulif.org.ua)

Визуализация разметки и интерфейс с пользователем

В результате семантической разметки фрагменты текста, связанные с терминами онтологии, помечаются открывающими и закрывающими тэгами с соответствующими именами. В дальнейшем при поиске по ключевым словам – терминам онтологии в тексте производится анализ только сами тэгов, и пользователю выводятся только те фрагменты, которые находятся между тэгами с соответствующими именами (или его подклассами или надклассом – по желанию пользователя).

В целом семантическая разметка визуализируется следующим образом: текст, находящийся между любыми двумя тэгами разметки, выводится на экран синим цветом. Если необходимо, то визуализироваться может не вся разметка, а только та, которая связана с заданным пользователем набором тэгов. Такой набор может задаваться явно (путем перечисления или выбора в списке тэгов) или неявно, через логические операции и отношения онтологии ПрО (например, все выбранные термины и их подклассы).

Система нормализации терминологии

Чтобы интегрировать различные виды естественно-языковых описаний и обрабатывать синонимию, необходимо:

1) в онтологии явным образом хранить сведения о том, что термин Т1 является синонимом термина Т2 (или,

например, его переводом на другой естественный язык);

2) разработать методы и средства трансформации в тексте всех упоминаний термина Т2 в термин Т1. Эта задача сводится к замене в тексте словоформ одного слова словоформами другого слова.

Замена словоформ одного слова словоформами другого слова выполняется в три этапа.

1. Определение морфологических характеристик словоформы. Такие характеристики находятся в морфологическом словаре.

2. Замена словоформ одного слова словоформами другого слова.

3. Проверка и исправление согласования с прилагательными и глаголами в роде, числе и падеже.

Если на первом этапе найдено более одного варианта, то имеет место случай омонимии: в языкознании омоним – это слово, совпадающее с другим по звучанию, но полностью расходящееся с ним по значению, а также по системе форм или по составу гнезда, например, "течь" и "течь2", "косить1" и "косить2". || прил. омонимический, -ая, -ое и омонимичный, -ая, -ое.

Для распознавания морфологической омонимии предлагается алгоритм упрощенного синтаксического анализа. В основе алгоритма лежит использование семантических характеристик слова, имеющихся в морфологическом словаре (одушевленность для существительных и переходность для глаголов).

Распознавание случая омонимии начинается с поиска ближайшего к существительному глагола (причастия/деепричастия) и определения его переходности. Непереходность глагола означает, что в предложении (причастном обороте, деепричастном обороте, обособленном определении, и т. п.) не может быть существительного в винительном падеже.

Например, в предложении «Коли ви не очікуєте податкового інспектора, він приходить саме до вас» форма “інспектора” соответствует родительному и винительному падежам, но глагол “очікувать” является непереходным, поэтому необходимо выбрать родительный падеж.

Очень часто для распознавания омонимии достаточно информации об одушевленности/неодушевленности. Рассмотрим пример “Рішення про перенiс строків подання податкової декларації приймає керівник місцевої адміністрації”. Здесь форма “Рішення” соответствует именительному, винительному и родительному падежу. Но слово “керівник” соответствует только именительному падежу, поэтому для слова «рішення» остается родительный и винительный падеж, а поскольку глагол “приймає” является переходным, то остается только винительный.

Формирование запроса по таксономии терминов ПрО с возможностью включения подклассов и надклассов

Например, если в запросе использован термин «корова», то запрос дополняется его надклассом «тварина» и его подклассами. Такой расширенный запрос позволит обнаружить в документе все упоминания о более общих и более частых случаях использования введенного ключевого слова. Разработанная система нормализации слов позволит обрабатывать слова, встречающиеся в документе в различных падежах, временах и т.п. Например, если в тексте встретилось «обробка телят», то пользователю этот фрагмент будет выделен по запросу «годуваня корів».

Семантическая обработка полученного запроса с учетом знаний о ПрО, содержащихся в онтологии, позволяет использовать слова-синонимы и слова, связанные с исходным термином в рамках ПрО (например, «корова» и «молоко»).

Трансформация ЕЯ-запроса в набор ключевых слов, соответствующих терминам онтологии (с использованием лексической онтологии) позволяет обрабатывать различные словоформы и в запросе, и в тексте закона.

Поиск в ЕЯ-текстах фрагментов, релевантных запросу пользователя, ведет-

ся на основе таких лексических знаний о ПрО, формализованных в виде лексической онтологии.

Программная реализация

Разработанная информационная система обеспечивает поиск в естественно-языковых документах, представляющих собой законы и нормативные акты, связанные с решаемой пользователем задачей (поиск релевантного набора документов находится вне рамок данного исследования) тех фрагментов текста, которые непосредственно связаны с конкретной информационной потребностью пользователя. Онтология ПрО обеспечивает знания о связях между терминами ПрО и позволяет заменять вводимые пользователем слова.

Рассмотрим работу системы пошагово на примере поиска нужной пользователю информации в Налоговом кодексе Украины.

Предположим, пользователя интересуют фрагменты этого документа, связанные с налогообложением транспортных средств. При этом будем считать, что пользователь знает, в каком документе находятся интересующие его сведения, но недостаточно разбирается в законодательстве, чтобы четко сформулировать свой запрос в соответствующих терминах.

Часть таких знаний отображены в лингвистической БЗ, позволяющей заменять термины на их синонимы, более привычные пользователю (и, соответственно, более понятные ему).

Сведения о структуре и иерархии понятий ПрО зафиксированы в виде построенной ранее онтологии. Например, с помощью такой онтологии можно получить информацию о том, что автомобиль является подклассом транспортного средства, а легковой транспорт – транспорта. Это позволяет при необходимости конкретизировать или расширять запрос, если его первоначальный вариант не позволяет обнаружить нужную информацию.

На первом шаге пользователь выбирает документ, в контенте которого будет осуществляться поиск, например, «Податковий кодекс». Затем он вводит ключевое слово или словосочетание для поиска. Предположим, пользователь ввел термин «автомашина», которого нет в Налоговом кодексе. Поэтому такой запрос при обычном поиске не даст нужного результата. Однако в лингвистической БЗ предложенной системы содержится информация, что для этого слова существует синоним «автомобиль», который и используется в анализируемом документе. Поэтому система трансформирует запрос, заменяя в нем слово «автомашина» на «автомобиль», и такой запрос уже находит все упоминания в тексте нужного пользователю термина.

Однако предполагается, что пользователю легче будет воспринять информацию в более привычной для него терминологии, и для этого предлагается осуществить замену всех вхождений слова «автомобиль» на слово «автомашина» с учетом форм слова.

В открывшемся окне «Параметры замены» (рис. 1) вводится слово, которое нужно заменить, и слово для замены. В соответствующих таблицах появляются все словоформы.

Следует заметить, что используемый для такой замены модуль относительно автономен, использует знания о способах изменения слов в естественный языках и может быть использован в других приложениях, анализирующих естественно-языковый текст – например, при семантической разметке, расширенном поиске или при переводе и редактировании специализированных документов.

Возвращаемся в главное окно и выбираем пункт меню «заменить». В списке сделанных замен появляются все встретившиеся случаи омонимии. Выбираем один из них, и видим то место в тексте, где была сделана замена.

После выполнения замены все найденные в документе слова выделяются цветом (рис. 2).

Если необходимо совершить поиск по всей иерархии классов вверх и вниз от заданного слова, то пользователю нужно воспользоваться пунктом меню «Расширенный поиск».

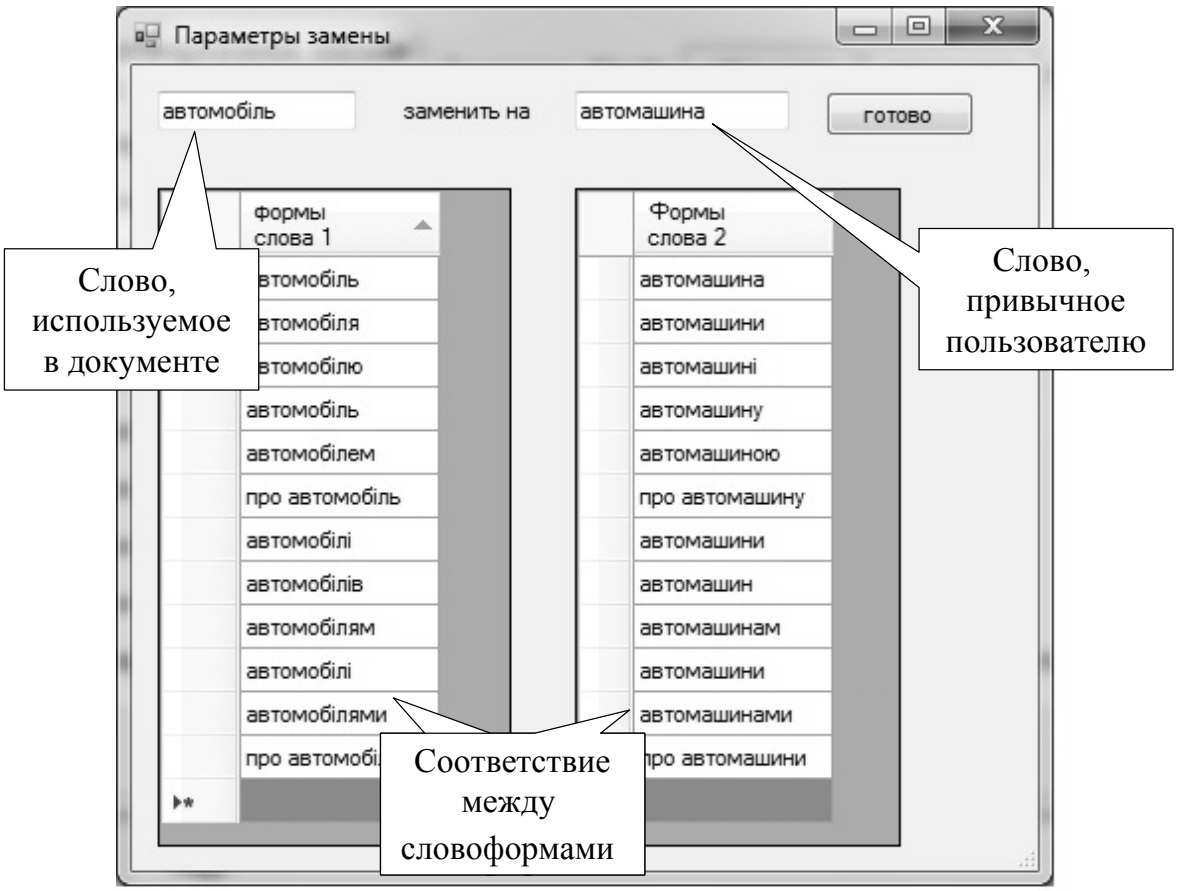

Рис. 1. Параметры замены слова с учетом словоформ

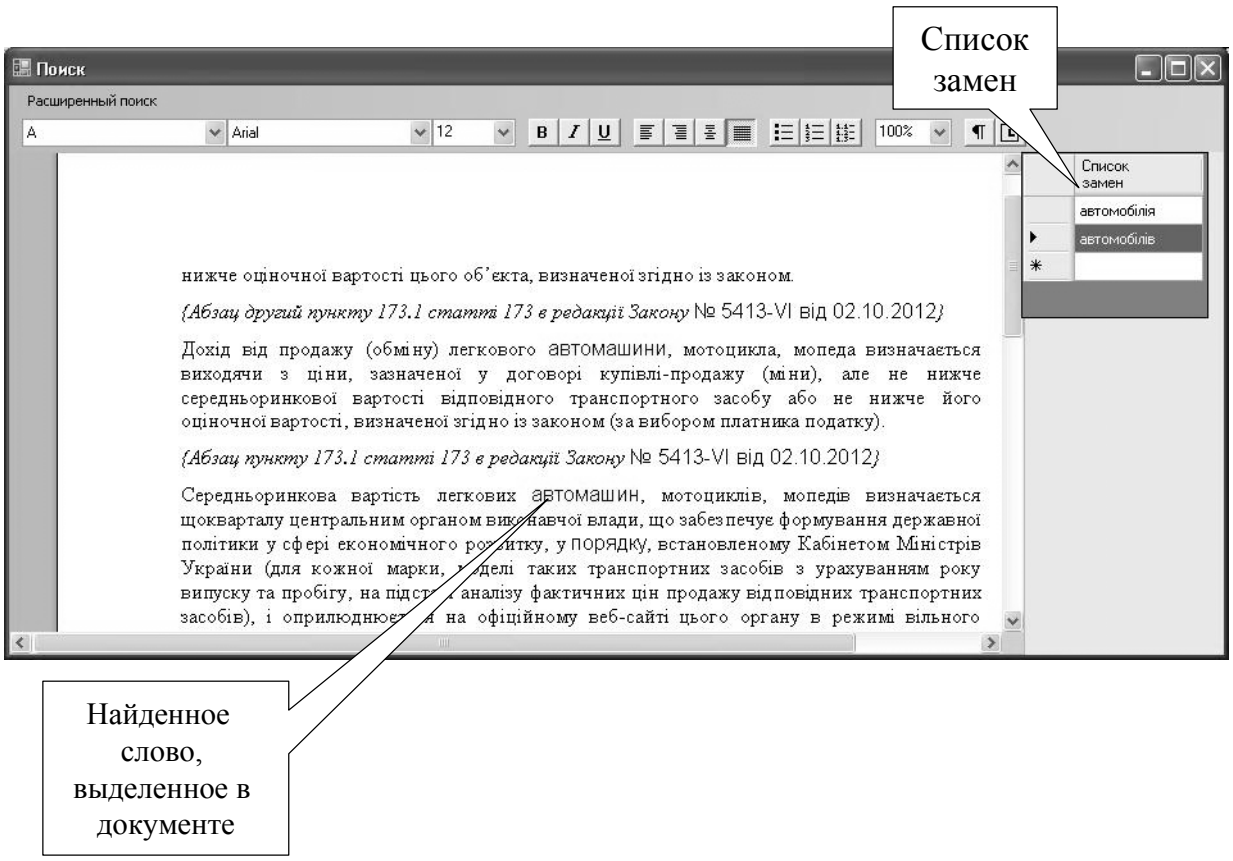

Рис. 2. Результаты поиска с учетом замены

Перспективы развития предложенной системы связаны с ее расширением на более широкие предметные области, характеризуемые естественно-языковыми документами.

Выводы

Описанный в работе основанный на онтологическом представлении знаний подход обеспечивает семантическую разметку естественно-языковых текстов терминами предметной области, содержащимися в соответствующей онтологии, что упрощает поиск пользователями нужной им информации.

1. *Лукашевич Н.В., Добров Б.В.* Проектирование лингвистических онтологий для информационных систем в широких предметных областях // Онтология проектирования. – 2015. – Т. 5, №1(15). – С. 47–69.
2. *Pedersen T., Patwardhan S., Michelizzi J.* WordNet: Similarity – measuring the relatedness of concepts // Proc. of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004. – P. 1024–1025.
3. *Рогушина Ю.В.* Разработка онтологической модели информационной потребности пользователя при семантическом поиске // Онтология проектирования. – 2014. – № 2 (12). – С. 61–82.
4. *Розробка* методів та засобів онтолого-лінгвістичного аналізу природно-мовних об'єктів // М.Г. Петренко, О.В. Палагін, В.Ю. Величко. С.Л. Кривий. – Київ: 2009. (препр., Інститут кібенетиким імені В.М. Глушкова НАН України). – 38 с.
5. *Лесько О., Рогушина Ю.* Использование специализированной лексической онтологии для автоматизации формирования онтологии предметной области по естественно-языковым текстам // Information Models of Knowledge. ITHEA, Kiev – Sofia, 2010. – P. 93–100.
6. *Добров Б.В., Лукашевич Н.В., Невзорова О.А., Федунов Б.Е.* Методы и средства автоматизированного проектирования прикладной онтологии // Известия РАН. Теория и системы управления. – М.: 2004. – № 2. – С. 58–68.
7. *Лесько О.Н., Рогушина Ю.В.* Использование онтологий для анализа семантики естественно-языковых текстов // Проблеми програмування. – 2009. – № 3. – С. 59–65.

1. *Lukashevich N.V., Dobrov B.V.* Design of linguistic ontologies for information systems in the broad subject areas // Ontology engineering. – 2015. – Vol. 5, N 1 (15). – P. 47–69. (in Russian).
2. *Pedersen T., Patwardhan S., Michelizzi J.* WordNet: Similarity – measuring the relatedness of concepts // Proc. of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), 2004. – P. 1024–1025.
3. *Rogushina J.V.* Design of the ontological model of user information need in semantic search // Ontology of design. – 2014. – № 2 (12). – P. 61–82 (in Russian).
4. *Design* of methods and means of ontological-linguistic analysis for natural language objects / M.G. Petrenko O.V. Palagin, V.Y. Velichko, S.L. Kriviy. – Kiev: 2009. (Preprint, Glushkov Institute of cybernetics). – 38 p. (in Ukrainian).
5. *Lesko O., Rogushina J.* Using of specialized lexical ontology for the automation forming of ontology of natural language texts // Information Models of Knowledge. ITHEA, Kiev – Sofia, 2010. – P. 93–100 (in Russian).
6. *Dobrov B.V., Lukashevich N.V., Nevzorova O.A., Fedunov B.E.* Methods and tools for automated design of applied ontology // Proceedings of RAN. Theory and control systems. – M.: 2004. – N 2. – P. 58–68 (in Russian).
7. *Lesko O.N., Rogushina Y.V.* Use of ontologies for analysis of natural language texts semantics // Problems of programming. – 2009. – N 3. – P. 59–65 (in Russian).

Получено 09.10.2015

Об авторах:

Лесько Ольга Николаевна
научный сотрудник Финансового управления НАН Украины.
Количество научных публикаций в украинских изданиях – 5.
Индекс Гирша – 1.
ORCID orcid.org/0000-0002-5584-3799,

Рогушина Юлия Витальевна,
кандидат физико-математических наук, старший научный сотрудник Института программных систем НАН Украины.
Количество научных публикаций в украинских изданиях – 100.
Количество научных публикаций в иностранных изданиях – 25.
Индекс Гирша – 10.
ORCID orcid.org/0000-0001-7958-2557.

Место работы авторов:

Финансовое управление НАН Украины.

Институт программных систем
НАН Украины,
03181, Киев-187,
Проспект Академика Глушкова, 40.
Тел.: (066) 550 1999.
E-mail: ladamandraka2010@gmail.com