УДК 004.41:47: 347.77

Е.А. ОРОБИНСКАЯ, аспирант, НТУ "ХПИ", Харьков, университет им. Люмьер Лион-2, Франция,
О.И. КОРОЛЬ, аспирант НТУ "ХПИ", Харьков, университет им. Люмьер Лион-2, Франция,
Н.В. ШАРОНОВА, д-р техн. наук, проф., НТУ "ХПИ", Харьков

ЯЗЫКОВАЯ КОМПЕТЕНЦИЯ ИНФОРМАЦИОННЫХ СИСТЕМ

В работе обсуждается стратегия автоматизированного построения онтологии на основе анализа патентно-конъюктурной информации; предлагается подход, основанный на использовании синтаксем русского языка, позволяющий обнаруживать в специализированных текстах термины данной предметной области. Библиогр.: 13 назв.

Ключевые слова: автоматизированное построение онтологии, синтаксема, информационная система.

Постановка проблемы. Предпосылкой появления прорывных технологий в области автоматической обработки текстов авторы считают поиск решений на основе методов системного анализа, как самого объекта исследования (текста), так и поставленных прикладных задач, которые должны быть решены в результате такой обработки. Иными словами, для семантического анализа текстовой информации требуется онтологический подход, и это, в свою очередь, означает, что для того чтобы обнаружить в тексте требуемую информацию, информационная система (ИС) сама должна обладать достаточным объемом лингвистических знаний.

Спектр же "интеллектуальных" задач, решение которых можно "поручить" таким ИС, чрезвычайно широк: от машинного перевода до формирования полноценных баз знаний, трактуемых сегодня как онтологии предметных областей. Непосредственно перед авторами стоит задача разработать систему обработки патентно-конъюнктурной информации (ПКИ) на основе анализа патентной документации, как правило, представляющей собой малоструктурированные тексты.

Анализ литературы свидельсвует о том, что в настоящее время активно ведется разработка программных средств, позволяющих автоматизировать процессы обработки информации [1, 2]. В [3] предложена классификация сущесвующих программных средств по их назначению, а в [1, 2, 4, 5] представлен систематизированный анализ методов, позволяющих обнаруживать и извлекать из текста конструктивные элементы для построения ИС или расширения уже существующей. Работы [6, 7, 8] доказывают, что именно использование

лингвистических методов позволяет существенно улучшить качестово автоматического анализа текста.

Цель работы – разработка лингвистического обеспечения для создания интеллектуальной системы автоматизированного построения онтологии на основе анализа и обработки текстовой патентно-конъюнктурной информации.

Текст как динамическая система. В общепринятом смысле под системой понимается множество взаимосвязанных элементов, обособленное от среды и взаимодействующее с ней, как целое. Несложно видеть, что патентная конъюнктурная информация, как и любая другая прикладная область, действительно представляет собой специальное множество с эмерджентными свойствами, обладающее структурной, функциональной и динамической организацией [9, 10].

Структурная организованность текста наблюдается со всей очевидностью: слова, предложения, абзацы, разделы и т.д. Число конструктивных элементов текста конечно и называется *словарем*. Слова формируют *переменные комбинации* по синтаксическим *правилам*, число которых также ограничено. С другой стороны, огромное множество конкретных способов и аспектов описания любой конкретной ситуации, кажется, могло бы существенно затруднить возможности естественнонаучного моделирования текста.

В рамках любой прикладной естественнонаучной дисциплины математическое моделирование оказывается эффективным благодаря выбору *ограниченного (конечного) числа степеней свободы* изучаемого явления, т.к. большинство степеней свободы динамической системы, на самом деле, являются связанными. В языке наблюдается та же всеобщая взаимосвязь слов, которая отражает иерархичность объектов действительности, обозначаемых этими словами – лексико-семантическими группами, связанными таксономическими отношениями. Для обозначения любого предмета можно последовательно указать все более общие понятия. Таким образом, мы неизбежно приходим к самому общему понятию "вещь" (объект, субстанция), являющемуся уже категорией познания, а не языка.

Патентная документация систематизируется от более общих к более узким тематическим и проблемным рубрикам в соответствии с международной классификацией изобретений (МКИ), что облегчает поиск требуемой информации. Проиллюстрировать сказанное можно на примере классификации, выполненной на основе патента на письменный компьютерный стол (http://base.uipv.org). Формула (1) демонстрирует таксономические отношения системы классификации описываемого

объекта [3, 4]. Здесь символом X обозначены предметные переменные, которые представляют конктретные концепты онтологии ПКИ:

$$\begin{cases} X_{1112147221}^{21/00} = X_{11121472}^{B} \wedge X_{1112147}^{47} \wedge X_{11121}^{A} \wedge X_{1112}^{Y} \wedge X_{111}^{H} \wedge X_{11}^{ИПМ} \wedge X_{1}^{ПС} \wedge X_{0}^{ИС}, \\ X_{1112147221}^{21/00} \wedge X_{1112147221}^{\overline{21/00}} = 0, \end{cases} \quad (1)$$

где $X_{1112147221}^{21/00}$ – компьютерный стол; $X_{11121472}^{B}$ – мебель; $X_{1112147}^{47}$ – артефакт; X_{11121}^{A} – жизненные потребности человека; X_{1112}^{Y} – устройство; X_{111}^{H} – новое устройство; $X_{11}^{ИПМ}$ – изобретение; $X_{1}^{ПС}$ – промышленная собственность; $X_{0}^{ИС}$ – интеллектуальная собственность.

То же касается описания свойств объектов и отношений между ними. Это значит, что и при вербальном описании ситуации наблюдается та же фактическая возможность существенно ограничить множество ее составляющих и свести все способы описания ситуации к упрощенной универсальной модели. Более того, очевидно, что ВСЕ возможные составляющие, не только не могут быть использованы в описании ситуации, но и не нужны, т.к. лишь затрудняют ее восприятие. Поэтому для распознавания некоторой ситуации достаточно отобрать лишь те составляющие, которые являются значимыми *с точки зрения эксперта-разработчика модели.*

Таким образом, мы видим, что текст соответствует понятию системы и может быть интерпретирован как структурно-функциональная, знаковая *модель* внешней ситуации.

Семантический анализ на основе синтаксем. Далее следует выявить, что именно нужно рассматривать в качестве элементарных структурных единиц текста, которые должны служить основой эффективного семантического анализа, определить их функциональные возможности и роли.

Согласно работам Г.А. Золотовой [13], синтаксический строй речи (текста) организуется регулярными комбинациями "элементарных" единиц, далее неделимых *на синтаксическом* уровне, из которых строятся все другие более сложные (речевые или текстовые) конструкции. В качестве такой единицы выдвинуто понятие *синтаксемы*. Синтаксемой, по Золотовой Г.А., называется минимальная далее неделимая семантико-синтаксическая единица языка, выступающая одновременно и как носитель элементарного смысла, и как конструктивный компонент с функциональностью, необходимой и

достаточной для построения более сложных синтаксических конструкций.

Признаками синтаксемы служат: 1) категориальное значение слова, от которого она образована; 2) конкретная морфологическая форма; 3) функциональность, вытекающая из двух первых признаков, как способность реализовываться в определенных позициях (речи, теста), как возможная роль в построении коммуникативной единицы.

Золотова Г.А. различает три функциональных роли: 1) возможность самостоятельного независимого существования синтаксемы; 2) использование синтаксемы в качестве компонента предложения; 3) использование синтаксемы в качестве компонента словосочетания, так называемое присловное употребление.

Функциональные свойства синтаксем служат основой для разграничения трех возможных типов, называемых, соответственно, *свободными* – А (или обладающими полным функциональным репертуаром), *обусловленными* – Б (способными выполнять функции 2, 3, крайне изредка –1), *связанными* – В (выступающими только в роли 3). Очевидно, что функциональные свойства синтаксем имеют транзитивный характер.

Рассматривая текст как динамическую систему, мы имеем возможность рассматривать процесс построения онтологии как целенаправленную операционную деятельность в своих пределах (т.е. в пределах данной системы), организованную для решения задач содержательного наполнения элементов онтологии. Само же понятие онтологии в терминах онтологического инжиниринга определено следующим образом [6]:

$$O = (C, \leq c, R, \sigma_R, \leq R, A, \sigma_A, T), \qquad (2)$$

где *C, R, A, T* – являются несвязанными множествами, чьи элементы называются идентификаторами концептов, отношений, атрибутов и типов данных соответственно; $\leq c$ – полусвязанная таксономия концептов с общим элементом самого верхнего уровня $root_C$; функция σ_R: $R \rightarrow C^+$, называемая признаком отношения *(relation signature);* $\leq_R$ on *R*, иерархия отношений где, $r_1 \leq_R r_2$ подразумевает $|\sigma_R(r_1)| = |\sigma_R(r_2)|$ и $\pi_i(\sigma_R(r_1)) \leq c \pi_i \sigma_R(r_2)$ для каждого $i \leq |\sigma_R(r_1)|$; функция σ_A: $A \rightarrow C \times T$, называемая *признаком атрибута (attribute signature);* множество *типов данных T*, таких как строки, целые числа и т.д.

В первом приближении можно согласиться с порядком, предложенным Симиано и др. [6] для построения онтологии на основе текста, известным как *layer cake technology*:

– сначала определяются термины-кандидаты, слова характеризуемые, как специфичные в нашей области ПКИ;

– затем найденные термины объединяются в семантически близкие группы (кластеры) на основании сравнения их атрибутов. Сформированным кластерам присваивается общая метка, называемая концептом;

– найденные концепты упорядочиваются в таксономические структуры;

– определяются ассоциативные связи между концептами;

– оформляются правила построения новых концептов.

Задача обнаружения терминов-кандидатов довольно успешно решается сегодня многочисленными статистическими методами [8]. Задачи их группировки также отчасти разрешимы этими методами (на основе анализа частот совместного появления слов на некотором ограниченном расстоянии). Но проблема обнаружения связей между понятиями не может быть удовлетворительно решена без привлечения лингвистических знаний.

Для обеспечения языковой компетентности, достаточной для самообучения и решения конечной задачи, т.е. построения патентной онтологии на базе текста, ИС сама должна обладать знаниями соответствующего порядка – общими (языковыми) и специальными (относящимися к конкретной предметной области). Такая ИС должна, по сути, объединять в себе две онтологии: общую онтологию языка (русского) и базовую (стартовую) онтологию ПКИ.

Проиллюстрируем сказанное.

Так, поссесив (т.е. отношения принадлежности, владения) в русском зыке может быть выражен следующими грамматическими формами: И.п., Р.п., у + Р.п., Т.п. (*именительный падеж, родительный падеж*, *предлог «у» плюс родительный падеж, творительный падеж*). Очевидно, что функции каждого падежа в русском языке разнообразны. Возможность «выхода» именно на отношения принадлежности обеспечивается указанием на категории слов, в сочетании с которыми данные падежи выражают отношения принадлежности. Так, например, глаголы {*обладать*, *обзаводиться, владеть*} в сочетании с творительным падежом существительного определяют именно эти отношения.

Надо признать, что создание ИС, обладающих таким анализатором, требуют больших усилий и затрат. Именно поэтому до последнего времени большинство инженерных решений для задач текстового анализа ограничивались статистическими методами, также приносящими положительные результаты. Но статистические методы, какими бы утонченными они не были, имеют заведомый предел своей точности и не

могут быть причисленными к интеллектуальным, оперирующим со смыслом текста.

Например, пусть имеется следующий набор потенциальных терминов-элементов для онтологии, содержащей ПКИ: "компания, мебель, представлять, жесткая конструкция, стол". Без дополнительной синтаксической информации невозможно определить точное содержание ситуации, а значит, и функцию каждого элемента: идет ли речь о возможном *представлении (описании)* стола как разновидности мебели) или о *действиях* какой-то компании, которая *представляет*, т.е. *предлагает* мебель в виде стола). Порядок слов в русском языке далеко не всегда может служить признаком выполняемой функции.

Все становится на свои места, если с перечнем слов мы получим дополнительную синтаксическую информацию об актантах глагола "представлять", например, о падеже актанта "стол":

– либо это (представлять + Т.п.), и тогда это указание на вероятное состояние: мебель представлена столом жесткой конструкции (субъект высказывания имеет модальную модификацию);

– либо это (представлять + И.п.), и тогда это указание на вероятное действие (так называемый потенсив), т.е. стол представляет мебель жесткой конструкции.

Поэтому тенденция современных исследований направлена именно на внедрение в ИС лексических знаний.

Заключение. Анализ текстовой патентно-конъюнктурной информации и извлечение из полнотекстовых документов релевантных данных остается актуальной задачей инженерии знаний в целом, и онтологического инжиниринга в частности. Качественное расширение возможностей ИС возможно при условии внедрения в них модулей, способных извлекать характеристики концептов на основе лингвистического анализа. Одним из возможных способов решения этой задачи является использование лингвистических шаблонов. Авторы предлагают общий подход на основе рассмотрения синтаксем русского языка. Поскольку каждая синтаксема описывается конечным детерминированным множеством признаков, такой подход является не только возможным, но и предпочтительным, поскольку он обеспечивает однозначное определение свойств концептов создаваемой онтологии. Трудоемкость задачи окупается качеством получаемых результатов. В следующей статье авторы планируют привести пример, иллюстрирующий работу лингвистического модуля разрабатываемой ИС для ПКИ и описать предлагаемый Фреймворк ИС.

Список литературы: 1. AI3’s Inaugural State of Tooling for Semantic Technologies / Adaptive Information Adaptive Innovation Adaptive Infrastructure // URL

http://www.mkbergman.com/991/the-state-of-tooling-for-semantic-technologies. **2.** *Ермаков А.Е.* Автоматизация онтологического инжиринга в системах извлечения знаний из текста **/** *А. Е. Ермаков* **//** Труды Международной конференции Диалог'2008. – Москва, Наука, 2008. – С. 136-140. **3.** *Corcho O.* Methodologies, tools, and languages for building ontologies. Where is their meeting point? / *O. Corcho, M. Fernandez-Lopez, A. Gomez-Perez* // Data & Knowledge Engineering. – 2003. – 46.– P. 41-64. **4***. Buitelaar P.* Ontology Learning from Texts: An Overview / *P. Buitelaar, P. Cimiano, B. Magnini* // In Ontology Learning from Text: Methods, Evaluation and Applications, 2005. – Vol. 123. – P. 234-265. **5.** *Simperl E.* Achieving Maturity: the State of Practice in Ontology Engineering **/** *E. Simperl, M. Mocho //* In International Journal of Computer Science and Applications, Technomathematics Research Foundation. – 2010. – Vol. 7. – № 1. – P. 45-65. **6.** *Makki J.* Semi Automatic Ontology Instantiation in the domain of Risk Management / *J. Makki, A.-M. Alquier, V. Prince //* In IFIP, Advances in Information and Communication Technology. – 2008. – Vol. 288. – P. 254. **7.** *Builеaar P.* Topic extraction from scientific literature for competency management / *P. Builеaar, T. Eigner* // In The 7th International Semantic Web Conference PICKME 2008, 27 octobre Karlsruhe, Germany. – P. 55-67. **8.** *Zhou L.* Ontology Learning: State of the Art and Open Issues / *L. Zhou* // Information Technology and Management. – 2007. – 8 (3). – P. 241-252. **9.** *Бондаренко М.Ф.* Теория интеллекта / *М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко*. – Харьков: Компания СМИТ, 2006. – 576 с. **10.** *Шаронова Н.В.* Автоматизированные информационные библиотечные системы: задачи обработки информации: монография, НУА / *Н.В. Шаронова, Н.Ф. Хайрова*. – Х.: 2003. – 120 с. **11.** *Король О.И.* Интеллектуальная обработка данных при формировании патентно-конъюнктурных баз знаний: Науково-технічний журнал "Біоніка інтелекту" / *О.И. Король, Н.В. Шаронова*. – Х.: Компания СМИТ. – 2012. – № 1 (78). – С. 12-16. **12.** *Король О.І.* Представлення й класифікація неструктурованих патентно-кон'юнктурних даних / *О.І. Король* // "Системи обробки інформації". – Харків. – 2011. – № 8 – С. 54-58. **13.** *Золотова Г.А.* Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса // *Г.А. Золотова.* – Едиториал УРСС, 2011. – 439 с.

УДК 004.41:47; 347.77

Мовна компетенція інформаційних систем / Оробінська О.О., Король О.І., Шаронова Н.В. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2012. – № 62 (968). – С. 148 – 154.

В роботі обговорюється стратегія автоматизованої побудови онтології на основі аналізу текстової патентно-кон'юнктурної інформації; запропоновано підхід, оснований на використанні синтаксем російської мови, що дозволяє знаходити у спеціалізованих текстах терміни даної предметної області. Бібліогр.: 13 назв.

Ключові слова: автоматизована побудова онтології, синтаксема, інформаційна система.

UDC 651.326

Technologi of reconfigurable computing / Orobinska O.O., Korol O.I., Sharonova N.V. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2012. – №. 62 (968). – P. 148 – 154

In this paper the strategy of automated ontology building is discussed for patent-conjuncture research. The method based on the utilization of Russian syntaxems is proposed. This method allows finding the terms of domain in the specialized texts. Refs.: 13 titles.

Keywords: automated ontology construction, syntaxem, informational system.

Поступила в редакцию 27.07.2012