ПОСТРОЕНИЕ ОНТОЛОГИЙ ДЛЯ СИСТЕМ ОБРАБОТКИ ПАТЕНТНО-КОНЪЮНКТУРНОЙ ИНФОРМАЦИИ

Оробинская Е. А., Король О. И.
Национальный технический университет НТУ «ХПИ» и Университет им. Люмьер Лион-2, Франция

Шаронова Н. В.
Национальный технический университет НТУ «ХПИ», г. Харьков, ул. Фрунзе, 21, тел. 057 707 64 60, e-mail: nvsharonova@mail.ru

В работе обсуждается стратегия автоматизированного построения онтологии; предлагается подход, основанный на использовании синтаксем русского языка, позволяющий обнаруживать в специализированных текстах термины данной предметной области. Рассмотрена задача обработки патентно-конъюнктурной информации (ПКИ) на основе анализа патентной документации, как правило, представляющей собой малоструктурированные тексты.

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

В общепринятом смысле под системой понимается множество взаимосвязанных элементов, обособленное от среды и взаимодействующее с ней, как целое. Патентная конъюнктурная информация, как и любая другая прикладная область, представляет собой специальное множество с эмерджентными свойствами, обладающее структурной, функциональной и динамической организацией [3, 4].

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. Для распознавания некоторой ситуации достаточно отобрать лишь те составляющие, которые являются значимыми с точки зрения эксперта-разработчика модели. Таким образом, текст соответствует понятию системы и может быть интерпретирован как структурно-функциональная, знаковая модель внешней ситуации.

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 [1]:

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

где C, R, A, T – являются несвязанными множествами, чьи элементы называются идентификаторами концептов, отношений, атрибутов и типов данных соответственно; ≤c – полусвязанная таксономия (semi-upper lattice) концептов с общим элементом самого верхнего уровня rootC; функция σR:R→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

σA: A→C×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. [1] для построения онтологии на основе текста, известным как layer cake technology:

- сначала определяются термины-кандидаты, – слова характеризуемые как специфичные в данной 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 [2]. 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.

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. 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.

Список литературы

1. *Buitelaar P.* Ontology Learning from Texts: An Overview. / *Buitelaar P., Cimiano P., Magnini B.* In Ontology Learning from Text: Methods, Evaluation and Applications, 2005, Vol. 123, Eds. IOS Press. P. 634.

2. *Zhou, L.* Ontology Learning: State of the Art and Open Issues/ Zhou, L. Information Technology and Management, 2007, 8(3), p. 241-252.

3. *Бондаренко М.Ф.* Теория интеллекта [Текст] : учеб./ *М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко*. – Харьков: Компания СМИТ, 2006. - 576 с.

4. *Шаронова Н.В*. Автоматизированные информационные библиотечные системы: задачи обработки информации [Текст]: монография, НУА / *Н.В. Шаронова, Н.Ф. Хайрова*; – Харьков 2003, – 120 с.