УДК 004.822:004.89

О ПРИМЕНЕНИИ ПАТТЕРНОВ ОНТОЛОГИЧЕСКОГО ПРОЕКТИРОВАНИЯ ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ НАУЧНЫХ ТЕКСТОВ

Боровикова Олеся Игнатьевна
М.н.с., e-mail: olesya@iis.nsk.su
Загорулько Юрий Алексеевич
К.т.н., зав. лабораторией, e-mail: zagor@iis.nsk.su
Кононенко Ирина Семеновна
Программист, e-mail: irina_k@cn.ru
Институт систем информатики им. А.П. Ершова СО РАН,
630090 г. Новосибирск, проспект Академика Лаврентьева, 6

Аннотация. В статье описан подход к автоматизированному извлечению информации из научных текстов на основе паттернов онтологического проектирования. Такие паттерны предназначены для описания решения типовых проблем, возникающих при разработке онтологий, и могут служить для представления как структурных, так и содержательных аспектов онтологии. Рассматриваются вопросы применения паттернов для решения задачи пополнения и построения онтологий с использованием информации о структуре онтологии и жанровых особенностях научных текстов. Представлены описания лексико-синтаксических паттернов, задающих отображение языковых структур в онтологические.

Ключевые слова: онтология, разработка онтологии, паттерны онтологического проектирования, извлечение информации, пополнение онтологии.

Цитирование: Боровикова О.И., Загорулько Ю.А., Кононенко И.С. О применении паттернов онтологического проектирования для извлечения информации из научных текстов // Информационные и математические технологии в науке и управлении. 2018. № 4 (12). С. 18–29. DOI: 10.25729/2413-0133-2018-4-02

Введение. Эффективность использования научных информационных интернет-ресурсов напрямую связана с тем, насколько полно и систематизировано представлена в них тематическая информация. Однако сбор, накопление и извлечение такой информации все еще остается довольно трудоемкой задачей, решить которую можно только путем автоматизации средств сбора и извлечения релевантной информации из текстов, представленных в сети Интернет.

Важную роль для разработки и поддержки таких ресурсов играет онтология, которая служит как для формализации и систематизации различных видов знаний, данных и информационных ресурсов, организации доступа к ним, так и для организации сбора, обработки и извлечения информации, интегрируемой в информационное пространство ресурса. Таким образом, разработка онтологии научной предметной области (НПО) является довольно ответственным и в то же время сложным и трудоемким процессом. Для снижения трудоемкости и упорядочения процесса разработки онтологии уже более десяти лет применяются методы, базирующиеся на применении онтологических паттернов

проектирования (Ontology Design Patterns или ODP) [21]. ODP представляют собой документально зафиксированные описания проверенных на практике решений проблем онтологического моделирования и позволяют описывать как типовые, так и специфические проблемы, возникающие при разработке онтологий, а также предлагаемые разработчиками рекомендации и соглашения для их решения.

В зависимости от типа проблем, для решения которых предназначены паттерны, различают несколько типов паттернов [21]. В данной работе мы остановимся на тех из них, использование которых актуально для решения задачи извлечения информации. В статье рассматриваются структурные паттерны и паттерны содержания, на основе которых выполняется построение онтологии НПО, и лексико-синтаксические паттерны, с помощью которых осуществляется ее наполнение. Таким образом, задача извлечения информации из научных текстов сводится к задаче автоматизированного пополнения онтологии, уже имеющей первоначальную структуру и наполнение.

Предлагаемый подход развивает методы сбора и извлечения онтологической информации [6], разработанные в рамках технологии построения интеллектуальных научных интернет-ресурсов.

1. Извлечение информации для пополнения онтологии. В существующих работах по автоматизированному пополнению онтологий и тезаурусов в зависимости от используемых методов выделяется три подхода: лингвистический, статистический и базирующийся на методах машинного обучения.

Лингвистический подход к решению задачи автоматического пополнения онтологии состоит в использовании предложенной в работе [17] идеи о возможности автоматизации построения семантических связей на основе диагностических контекстов, представленных в виде лексико-синтаксических шаблонов (ЛСШ). Данный метод, известный как шаблоны Херст (Hearst patterns), предназначен для обработки неструктурированных текстов. Он широко применялся для извлечения родовидовых отношений и предполагает извлечение из коллекции документов упорядоченных пар слов, соответствующих множеству заранее составленных шаблонов. Подход M. Hearst использовался и совершенствовался многими другими исследователями, а также применялся для других языков.

В ряде работ предлагается формальный аппарат для записи лексических и лексико-синтаксических шаблонов. Так, в работе [13] формулируется XML-схема языка для формализации лексико-синтаксических шаблонов, используемых для пополнения онтологий. Система Alex [5] и являющийся ее развитием инструмент DigLex [9] предоставляют довольно гибкие средства описания слов и словосочетаний в виде шаблонов, которые используются затем для автоматического распознавания этих единиц в тексте. Они расширяют возможности традиционных лексикографических систем: язык описания шаблонов поддерживает использование альтернатив, ссылки на шаблоны, повторители, условия на контекст, дистантный контекст и т.п. Язык позволяет записывать правила не только для распознавания текстовых объектов, но и для определения их лексических и семантических атрибутов. Однако в языках Alex и DigLex нет встроенных средств для указания грамматических признаков распознаваемых лексических единиц и грамматического согласования нескольких единиц, необходимых для однозначного выделения языковых конструкций (например, именных групп). Этого последнего недостатка лишен

предложенный в работе [2] язык LSPL, позволяющий задавать грамматические свойства входящих в него элементов.

Разные авторы, например, [16], отмечают, что общей проблемой метода построения связей по тексту на основе лексико-синтаксических шаблонов является разреженность информации, и без использования внешних семантических ресурсов невозможно восстановить связь между близкими по значению словами в случае их отсутствия в одном фрагменте текста. Отмечается и тот недостаток, что из извлеченных отношений достаточно сложно сформировать полноценную иерархию связей.

В последнее десятилетие для пополнения онтологии получило развитие направление [14], использующее частично структурированные тексты: словари, энциклопедии и др., главным образом, англоязычные. При использовании словарей обычно в том или ином виде применяется метод шаблонов или его аналог. Шаблоны в данном случае опираются на жанровые особенности конкретного типа словаря – структуру словарной статьи и лексико-синтаксическую структуру словарной дефиниции. Так, при использовании словаря синонимов [11] или викисловаря [10, 16] авторы существенным образом опираются на такие элементы структуры словарной статьи, как лексикографические пометы, структура тегов и т.п. Извлечение информации из словарных определений затруднено неидеальным семантическим качеством самих словарных определений, которые даются в энциклопедиях и толковых словарях [14]. Второй проблемой является несовершенное качество автоматического понимания текстов на ЕЯ на современном уровне развития лингвистических технологий, что предполагает активное участие эксперта на этапе оценки полученных результатов. В связи с этим в некоторых работах (например, [8]) декларируется автоматическое извлечение пар терминов-кандидатов на родовидовую связь, чтобы получить не гарантированные, а лишь вероятные пары.

В работе [20] извлекаются таксономические и 10 нетаксономических отношений из глоссария тезауруса по изобразительному искусству и архитектуре, где фиксируются текстовые дефиниции терминов. Аннотирование дефиниций выполняется с помощью регулярных выражений, в которых традиционно используется лексическая и частеречная информация и добавлены синтаксические и семантические ограничения. Особенность этой работы не только в расширенных шаблонах, но и в использовании WordNet для разрешения лексической неоднозначности.

Статистический подход лежит в основе этапа формирования терминологической базы тезауруса или онтологии [3]. В работе [12] построение множества ассоциативных и иерархических отношений по тексту происходит путем формирования множества не типизированных отношений-кандидатов между дескрипторами тезауруса. Для нахождения связей синонимии поиск отношений производится в заранее составленной базе – словаре синонимов. Поиск иерархических и ассоциативных отношений осуществляется непосредственно в тексте на основе анализа совместной встречаемости слов, исходя из которого эксперт принимает дальнейшее решение о наличии связей между концептами тезауруса. Статистический подход к извлечению онтологической информации из научных текстов в [15] основан на т.н. индикаторном методе, в основе которого лежит обнаружение в тексте специфических подсказок в виде различного рода словесных клише (образцов, маркеров) типа: «в настоящей работе», «в работе рассматривается», «целью… является», «новый подход к», «предлагается использовать», «проведенное исследование» и т.п. Эти

клише являются индикаторами того или иного аспекта содержания текста. Основы индикаторного подхода к извлечению информации из текста были заложены еще в 70-е годы прошлого столетия. Достаточно детальное описание подхода представлено в [1]. В [15] проиллюстрирована возможность использования количественных характеристик L-граммного спектра для частичной автоматизации процедуры формирования и обогащения (путем варьирования) индикаторных словарей, фиксирующих подсказки о различных аспектах содержания научного текста.

Методы обучения для автоматического формирования понятий и связей между ними представлены в работах [16], [18] и др. В [18] реализуется подход на основе обучения, исходя из некоторого начального множества пар терминов для извлечения лексико-синтаксических шаблонов семантических отношений из текстов интернета. В работе [16] формируется семантическая сеть слов путем связывания отдельных лексических значений слов (а не понятий или синсетов). Комплекс программ включает в себя реализацию методов обнаружения групп синонимов и построения отношений между отдельными значениями слов, основанных на обучении без учителя, а также модуль расширения отношений, основанный на обучении с учителем.

2. Паттерны для автоматизированного пополнения онтологии. Сложность задачи сбора информации для научных информационных ресурсов определяется большим разнообразием видов извлекаемой информации и способов ее представления в сети Интернет. Информация для научных ресурсов пополняется из таких источников, как порталы знании, словарные и энциклопедические ресурсы, электронные библиотеки и журналы, сайты организаций, ассоциаций, проектов и конференций, новостные ленты, социальные научные сети, вики-ресурсы, реестры (каталоги) веб-сервисов и др. Эта информация может быть представлена как в виде интернет-страниц, имеющих различную структуру, так и в виде текстовых документов в различных форматах.

Из текстов этих источников необходимо извлекать/собирать информацию обо всех сущностях, которые описываются онтологией НПО, т.е. о проектах, организациях, персонах, конференциях, публикациях, разделах науки, методах, объектах и предметах исследований.

Первоначальное формирование онтологии НПО осуществляется на основе структурных паттернов и паттернов содержания онтологического проектирования.

2.1.Структурные паттерны и паттерны содержания онтологии НПО. *Структурные паттерны* либо фиксируют способы решения проблем, вызванных ограничениями выразительных возможностей языков описания онтологий, либо задают общую структуру и вид онтологии. Необходимость в использовании структурных логических паттернов для построения онтологии НПО вызвана проблемой отсутствия в языке OWL выразительных средств для представления сложных сущностей и конструкций, актуальных при построении онтологии, в частности, областей допустимых значений и многоместных отношений с атрибутами. Эти паттерны являются предметно-независимыми, на их основе могут строиться фрагменты онтологии, входящие в паттерны содержания.

Паттерны содержания задают способы представления типовых фрагментов онтологий, на основе которых могут строиться онтологии целого класса предметных областей. Применение паттернов содержания при создании онтологии НПО позволяет обеспечить возможность единообразного и согласованного представления используемых в

ней научных понятий и их свойств, на основе описания понятий и видов выполняемой научной деятельности, характерных для большинства научных предметных областей.

В рамках методологии построения онтологий НПО [7] разработан набор паттернов содержания для представления таких понятий, как *Объект исследования, Предмет исследования, Метод исследования, Раздел науки, Научный результат, Персона, Организация, Деятельность (Научная деятельность), Проект, Публикация и др.* При разработке этих паттернов для каждого из них был определен набор квалификационных вопросов, представляющих его содержание. С помощью этих вопросов был выявлен обязательный и факультативный состав онтологических элементов паттернов и описаны требования к ним, представленные в виде аксиом и ограничений.

Для каждого онтологического элемента, представленного паттерном содержания или структурным паттерном, разрабатывается набор лексико-синтаксических паттернов, который отражает разнообразие способов представления этого элемента в научных текстах.

2.2.Лексико-синтаксические паттерны онтологии НПО. *Лексико-синтаксические паттерны* применяются для автоматизации построения и пополнения онтологий на основе текстов на естественном языке и задают отображения языковых структур в онтологические структуры [19]. Элементами лексико-синтаксических паттернов могут быть группы слов и словосочетаний определенного языка, соответствующие онтологическим конструкциям, заданным как в языке описания онтологий, так и в структурных (логических) паттернах и паттернах содержания.

Идея этого типа паттернов основывается на концепции лексико-синтаксических шаблонов языковых конструкций для извлечения из текста языковых единиц [2, 4, 5, 17], которые, как показано в разделе 1, могут успешно применяться и для извлечения понятий и связей онтологии.

На основе лексико-синтаксических паттернов может осуществляться не только распознавание экземпляров классов понятий и выявление родовидовых связей между ними для построения и дополнения иерархии, но и выполняться извлечение свойств (атрибутов и связей), характерных для объектов определенного класса. Кроме того, данные паттерны могут включать элементы (жанровые маркеры), которые отражают жанровую специфику текста.

Жанр научной прозы относится к числу наиболее изученных как с формально-лингвистической, так и со статистической точки зрения. В [15] рассмотрено 12 аспектов содержания текстов жанра научной статьи, таких как «цель/задача исследования», «элементы новизны», «метод решения», «полученные результаты» и др., и путем просмотра их L-граммных характеристик выделено около 700 потенциально возможных лексических жанровых маркеров-индикаторов. Так, для аспекта «цель/задача исследования» выделены наиболее сильные индикаторы, продемонстрировавшие 100%-ю точность (*в\статье\рассматриваться, в\работе\рассматриваться, в\работе\обсуждаться, цель\этой\работы, данная\работа\посвящать* и др.), и более слабые индикаторы (*предлагаться, исследоваться, ставиться\задач, наше\исследование, в\своей\работе, в\рамках\проекта*). С целью повышения степени полноты и точности идентификации аспекта исходный словарь был расширен путем варьирования его элементов за счет синонимичных или условно синонимичных подстановок (*работа/статья/доклад/сообщение/исследование*; *рассматриваться/анализироваться/*

исследоваться), а также варьирования на уровне словообразования (*рассмотреть/рассматривать/рассматриваться, предлагаемый/предложенный, важный/важнейший/особо важный/особенно важный*). Для повышения эффективности распознавания в случае более слабых индикаторов требуется их включение в состав конструкций, учитывающих грамматические характеристики самих индикаторов и связанных с ними элементов, т.е. создание лексико-синтаксических шаблонов.

С помощью структурных паттернов определяется, какая онтологическая конструкция должна извлекаться из текста, например, значение домена или атрибут отношения, в то время как паттерны содержания определяют состав объектов и связей представленных в них понятий. Так, например, с помощью паттерна содержания, предназначенного для описания метода исследования, генерируется набор лексико-синтаксических паттернов для извлечения объектов и свойств класса *Метод исследования* (Рис. 1).

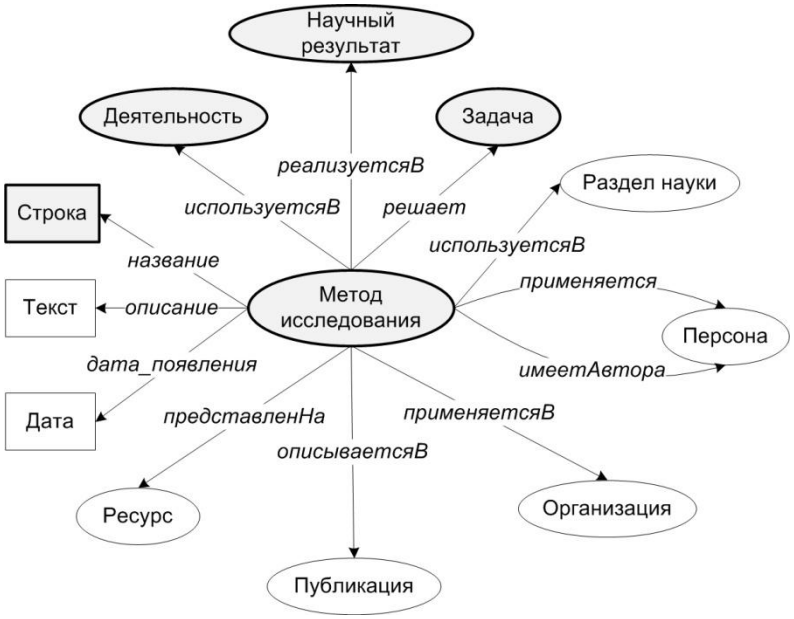

Рис. 1. Паттерн содержания для описания метода исследования

Рассмотрим примеры лексико-синтаксических паттернов, описанных с помощью модификации предложенного в работе [2] языка для извлечения информации из научно-технических текстов. Предлагаемая нами модификация языка, в частности, учитывает возможность разрыва в составе шаблона.

Следующие паттерны позволяют извлекать информацию о свойстве *название* описываемого в тексте экземпляра класса *Метод исследования* и используют предварительно заданные шаблоны именной группы (NP), глагольной лексемы, включающей личные формы глагола и причастия (VPa= V| Pa), и глагольной группы (VP и VPaP). PnP обозначает личное местоимение, точка обозначает разрыв, в квадратные скобки помещается факультативный элемент:

NMi = NP1< метод, c=acc> [.] VPa <называть> [.] NP2 <c=ins> <NP1.n = NP2.n>

NMj = VPaP <предлагать> [.] NP1< метод > [.] VPa <называть> [.] NP2 <c=ins> <NP1.n = NP2.n>

Эти шаблоны описывают, в частности, фразы: *Описывается метод, называемый методом согласования для обращения соответствующих интегральных уравнений; По аналогии с известным расчетным статическим методом предлагаемый метод назовем расчетным динамическим; В данной работе предлагается метод, называемый «элитным отбором» или «элитной стратегией»; Описываемый метод определения периода гаммы в шифре гаммирования по известному шифртексту B= b1,b2,…,bN назовем методом ББШ; Предлагаемый метод назовем методом двух групп; В работе предлагается метод, называемый Voxel Cone Tracing (VCT); Для анализа смесей веществ, содержащих нестабильные и, в том числе, взаимодействующие друг с другом компоненты, претерпевающие превращения в испарителе хроматографа, предлагается метод, называемый далее терюкинетическим; В качестве альтернативного способа предлагается метод, называемый криотерапией; Применяемый в работе метод называется методом магнетрона.*

Снятие ограничения c=ins расширяет класс распознаваемых конструкций: *В работе предлагается метод называемый иерархическая редукция (hierarchical reduction).*

Добавление следующего паттерна позволяет учесть достаточно частотные в научных статьях конструкции с местоименным анафорическим элементом, такие как *Предлагаемый метод, назовем его «методом перекоса»; Предлагаемый метод, назовем его «оценочное взвешенное пересечение»*:

NMk= VPaP <предлагать> [.] NP1< метод > [.] VPa <"называть"> [.] NP2 <PnP, c= acc > [.] NP3 < NP1.n = NP2.n, NP1.p = NP2.p>

Введение в состав паттерна глагольной группы на базе таких глаголов, как *участвовать, использоваться, применяться, развиваться, предлагаться, апробироваться*, позволит выявлять отношения *используетсяВДеятельности*. Именная группа *научное исследование, научная разработка, научный проект, научная программа* будет указывать на объект (экземпляр) класса *Деятельность*.

Таким образом, из текста конкретного жанра с помощью характерных для данного жанра шаблонов извлекаются целевые понятия и связи, зафиксированные в паттерне содержания.

Заключение. В статье рассмотрены вопросы использования паттернов онтологического проектирования для построения и пополнения онтологий.

Непосредственно для этих целей используются лексико-синтаксические паттерны, которые строятся на основе паттернов содержания. При этом для каждого фрагмента паттерна содержания, описывающего семантическую связь входящих в него понятий, разрабатывается отдельный набор лексико-синтаксических паттернов, каждый из которых отражает разнообразие способов представления этой связи в научных текстах.

Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований (грант № 16-07-00569) и Президиума СО РАН (Блок 36.1. Комплексной программы ФНИ СО РАН II.1).

СПИСОК ЛИТЕРАТУРЫ

1. Блюменау Д.И., Гендина Н.И., Добронравов И.С., Лахути Д.Г., Леонов В.П., Федоров Е.Б. Формализованное реферирование с использованием словесных клише (маркеров) // Научно-техническая информация. Сер.2. 1981. №2. С. 16–20.

2. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2007. М.: Издательский центр РГГУ. 2007. С. 70–75.
3. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды V Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL 2003. Санкт-Петербург. 2003. С. 201–210.
4. Ермаков А.Е., Плешко В.В., Митюнин В.А. RCO Pattern Extractor: компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Межд. научная конференция. Сборник трудов. Москва. 2003. С. 312–317.
5. Жигалов В.А., Жигалов Д.В., Жуков А.А., Кононенко И.С., Соколова Е.Г., Толдова С.Ю. Система Alex как средство для многоцелевой автоматизированной обработки текстов // Труды международного семинара Диалог'2002 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука. 2002. Т.2. С. 192–208.
6. Загорулько Ю.А., Боровикова О.И., Сидорова Е.А., Ахмадеева И.Р. Сбор онтологической информации для интеллектуальных научных Интернет-ресурсов // Системная информатика. 2014. № 3. С. 13–23.
7. Загорулько Ю.А., Загорулько Г.Б., Боровикова О.И. Технология создания тематических интеллектуальных научных интернет-ресурсов, базирующаяся на онтологии // Программная инженерия, 2016. № 2. С. 51–60.
8. Киселев Ю.А., Поршнев С.В., Мухин М.Ю. Метод извлечения родовидовых отношений между существительными из определений толковых словарей // Программная инженерия. 2015. № 10. С. 38–48.
9. Ковалев А.И., Сидорова Е.А. Инструмент разработки предметных словарей на основе лексических шаблонов DigLex // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–2015), 6 - 8 октября 2015 г., Новосибирск. Новосибирск: Институт математики им. С.Л. Соболева СО РАН. 2015. Т. 1. С. 123–130.
10. Крижановский А.А., Смирнов А.В. Подход к автоматизированному построению общецелевой лексической онтологии на основе данных Викисловаря // Известия РАН. Теория и системы управления. 2013. № 2. С. 53–63.
11. Оробинская Е.А. Метод автоматического построения онтологии предметной области на основе анализа лингвистических характеристик текстового корпуса // Интернет и современное общество (IMS-2012): тр. XV Всерос. объединенной конф. СПб. 2012. С. 209–212.
12. Панченко А. Технология автоматизированного построения информационно-поискового тезауруса. Режим доступа: http://it-claim.ru/Persons/Panchenko/article_thesauri.pdf.
13. Рабчевский Е.А. Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2009. Петрозаводск. 2009. С. 69–77.

14. Рубашкин В.Ш., Бочаров В.В., Пивоварова Л.М., Чуприн Б.Ю. Опыт автоматизированного пополнения онтологий с использованием машиночитаемых словарей // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). М.: Изд-во РГГУ. 2010. С. 413–418.
15. Саломатина Н.В., Гусев В.Д. Автоматизация формирования индикаторных словарей и возможности их использования // Труды межд. конференции Диалог-2006 «Компьютерная лингвистика и интеллектуальные технологии», Бекасово, 31мая − 4 июня 2006. Москва. "Наука". С. 121−125.
16. Усталов Д.А., Созыкин А.В. Комплекс программ автоматического построения семантической сети слов//Вестник Южно-уральского государственного университета. Сер. Вычислительная математика и информатика. 2017. Т. 6. № 2. С. 69–83.
17. Hearst M.A. Automatic Acquisition of Hyponyms from Large Text Corpora // In: Proceedings of the 14th International Conference on Computational Linguistics. 1992. Pp. 539–545.
18. Kozareva Z., Hovy E. Learning arguments and supertypes of semantic relations using recursive patterns. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2010. Pp. 1482–1491.
19. Maynard D., Funk A., Peters W. Using lexico-syntactic ontology design patterns for ontology creation and population. In Proc. of WOP2009 collocated with ISWC2009. V. 516. Pp. 39–52. CEUR-WS.org.
20. Navigli R., Velardi P. From Glossaries to Ontologies:Extracting Semantic Structure from Textual Definitions // Ontology learning and population: bridging the gap between text and knowledge / Buitelaar P., Cimiano P. (eds.) // Series: Frontiers in artificial intelligence and applications, vol. 167. Amsterdam; Washington. DC: IOS Press. 2008. Pp. 71–87.
21. Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web / Hitzler P., Gangemi A., Janowicz K., Krisnadhi A., Presutti V. (eds.) IOS Press/AKA. 2016.

UDK 004.822:004.89

ON APPLICATION OF ONTOLOGY DESIGN PATTERNS FOR EXTRACTION OF INFORMATION FROM SCIENTIFIC TEXTS

Olesya I. Borovikova
Junior Researcher, e-mail: olesya@iis.nsk.su
Yury A. Zagorulko
Dr., Head of Laboratory "Artificial Intelligence", e-mail: zagor@iis.nsk.su
Irina S. Kononenko
Programmer of Laboratory "Artificial Intelligence", e-mail: irina_k@cn.ru
A.P. Ershov Institute of Informatics Systems
Siberian Branch of the Russian Academy of Sciences
6, Acad. Lavrentjev pr., 630090, Novosibirsk, Russia

Abstract. The paper describes an approach to the automated extraction of information from scientific texts based on ontology design patterns. Such patterns are intended to describe the solution of typical problems arising in the development of ontologies, and can serve to represent both structural and semantic aspects of ontology. The questions of applying patterns to solve the problem of population and construction of ontologies with the use of information on the ontology structure and genre features of scientific texts are considered. The descriptions of the lexico-syntactic patterns are presented for the mapping of language constructions into ontological structures.

Keywords: ontology, ontology development, ontology design patterns, information extraction, ontology population.

References

1. Bljumenau D.I., Gendina N.I., Dobronravov I.S., Lahuti D.G., Leonov V.P., Fedorov E.B. Formalizovannoe referirovanie s ispolzovaniem slovesnykh klishe [Formalized Summarization by Using Verbal Clichés (markers)] // Nauchno-tehnicheskaja informatsija = Scientific and Technical Information. 1981. Ser. 2. № 5. Pp. 16–20. (in Russian)
2. Bolshakova E.I., Baeva N.V., Bordachenkova E.A., Vasilieva N.E., Morozov S.S. Leksiko-sintaksicheskie shablony v zadachah avtomaticheskoj obrabotki tekstov [Lexicosyntactic patterns for automatic text processing] // Kompjuternaja lingvistika i intellectual'nye technologii: Trudy Mezdunarodnoj konferentsii Dialog'2007. M.: Izdatelskij tsentr RGGU. 2007. Pp. 70–75. (in Russian)
3. Dobrov B.V., Loukachevitch N.V., Syromyatnukov S.V. Phormirovanie bazy terminologicheskikh slovosochetanij po tekstam predmetnoj oblasti [Automatic Detection of Text Entries for Information Retrieval Thesaurus] // Trudy V Vserossijskoj nauchnoj konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronye kollektsii» - RCDL 2003. St. Petersburg. 2003. Pp. 201–210. (in Russian)
4. Ermakov A.E., Pleshko V.V., Mityunin V.A. RCO Pattern Extractor: komponent vydelenija osobykh ob'ektov v tekste [RCO Pattern Extractor: program to extract special constitients from text] // Informatizatsija i informatsionnaja bezopasnost' pravookhranitel'nykh: XI Mezhd. Nauchnaja konferentsija. Sbornik trudov – Moscow, 2003. pp. 312–317. (in Russian)

5. ZhigalovV.A., Zhigalov D.V., Zhukov V.V., Kononenko I.S., Sokolova E.G., Toldova S.Yu. Sistema Alex kak sredstvo dlja mnogotselevoj avtomatizirovannoj obrabotki tekstov [ALEX - a system for multi-purpose automatized text processing] // Trudy mezdunarodnogo seminara Dialog'2002 «Kompjuternaja lingvistika i intellectual'nye technologii». M.: Nauka. 2002. V.2. Pp.192–208. (in Russian)
6. Zagorulko Yu.A., Borovikova O.I., Sidorova E.A., Ahmadeeva I.R Sbor ontologicheskoj informatsii dlja intellektual'nykh nauchnykh Internet-resursov [An ontological information collection for intelligent scientific internet resources] // Sistemnaja informatika = System Informatics. 2014. №3. Pp.13–23. (in Russian)
7. Zagorulko Yu.A., Zagorulko G.B., Borovikova O.I. Tekhnologija sozdanija tematicheskikh intellektual'nykh nauchnykh internet-resursov, bazirujushhajasja na ontologii [Technology for building subject-based intelligent scientific internet resources based on ontology] // Programmnaja inzhenerija = Software Engineering. 2016. № 2. Pp. 51–60. (in Russian)
8. Kiselev Yu.A., Porshnev S.V., Mukhin M.Yu. Metod izvlechenija rodovidovykh otnoshenij mezhdu sushchestvitel'nymi iz opredelenij tolkovykh slovarej [Method of Extracting Hyponym-Hypernym Relationships for Nouns from Definitions of Explanatory Dictionaries] // Programmnaja inzhenerija = Software Engineering. 2015. № 10. Pp. 38–48. (in Russian)
9. Kovalev A.I., Sidorova E.A. Instrument razrabotki predmetnykh slovarej na osnove leksicheskikh shablonov DigLex [Tool for developing subject dictionaries based on lexical templates DigLex] // Materialy Vserossijskoj konferentsii s mezhdunarodnym uchastiem «Znanija-Ontologii-Teorii» (Zont–2015). Novosibirsk: Institut matematiki im.S.L.Soboleva SO RAN. 2015. V.1. Pp. 123–130. (in Russian)
10. Krizhanovsky A.A., Smirnov A.V. Podkhod k avtomatizirovannomu postroeniyu obshchetselevoj leksicheskoj ontologii na osnove dannykh Vikislovarja [An approach to automated construction of a general-purpose lexical ontology based on wiktionary] // Izvestija RAN. Teorija i sistemy upravlenija = Theory and Control systems. 2013. № 2. Pp. 53–63. (in Russian)
11. Orobinska E.A. Metod avtomaticheskogo postroenija ontologii predmetnoj oblasti na osnove analiza lingvisticheskikh kharakterestik tekstovogo korpusa [Automatic Method Of Domain Ontology Construction based on Characteristics of Corpora POS-Analysis] // Internet i sovremennoe obshchestvo (IMS-2012): tr. XV Vseross. ob'edin. konf. S-Pb., 2012. Pp. 209–212. (in Russian)
12. Panchenko A. Tekhnologija avtomatizirovannogo postroenija informatsionno-poiskovogo tezaurusa [Technology of the automated thesaurus construction for Information Retrieval]. Accessed at: http://it-claim.ru/Persons/Panchenko/article_thesauri.pdf. (in Russian)
13. Rabchevsky E.A. Avtomaticheskoe postroenie ontologii na osnove lrksiko-sintaksicheskikh shablonov dlja informatsionnogo [Automatic ontology construction based on lexical-syntactic patterns for information retrieval] // Trudy XI Vserossijskoj nauchnoj konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronye kollektsii» - RCDL'2009. Petrozavodsk. 2009. Pp. 69–77. (in Russian).
14. Rubashkin V.Sh., Bocharov V.V., Pivovarova L.M., Chuprin B.Ju. Opyt avtomatizirovannogo popolnenija ontologij s ispolzovaniem mashinochitaemyh slovarej [The approach to ontology learning from machine-readable dictionaries] // Kompjuternaja lingvistika i intellectual'nye

technologii: Trudy Mezdunarodnoj konferentsii Dialog'2010. Iss. 9 (16). M.: Izdatelstvo RGGU. Pp. 413–418. (in Russian)

15. Salomatina N.V., Gusev V.D. Avtomatizatsija formirovanija indikatornykh slovarej i vozmozhnosti ikh ispolzovanija [Automation of CUE dictionaries formation and their applications] // Trudy mezhdunarodnoj konferentsii Dialog-2006 «Kompjuternaja lingvistika i intellectual'nye technologii». Moscow. Nauka. Pp. 121−125. (in Russian)
16. Ustalov D.A, Sozykin A.V. Komleks programm avtomaticheskogo postroenija semanticheskoj seti slov [A software system for automatic construction of a semantic word network] // Vestnik Yuzhno-ural'skogo gosudarstvennogo universiteta. Ser.: Vychislitel'naja matematika i informatika = Bulletin of South Ural State University. Computational Mathematics and Software Engineering. 2017. V. 6 № 2. Pp. 69–83. (in Russian)
17. Hearst M.A. Automatic Acquisition of Hyponyms from Large Text Corpora. //In: Proceed-ings of the 14th International Conference on Computational Linguistics. 1992. Pp. 539–545.
18. Kozareva Z., Hovy E. Learning arguments and supertypes of semantic relations using recursive patterns. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2010. Pp. 1482–1491.
19. Maynard D., Funk A., Peters W. Using lexico-syntactic ontology design patterns for ontology creation and population. In Proc. of WOP2009 collocated with ISWC2009. vol. 516. Pp. 39–52. CEUR-WS.org.
20. Navigli R., Velardi P. From Glossaries to Ontologies: Extracting Semantic Structure from Tex-tual Definitions // Ontology learning and population: bridging the gap between text and knowledge / Buitelaar P., Cimiano P. (eds.) Series: Frontiers in artificial intelligence and applications. vol. 167. Amsterdam; Washington, DC: IOS Press. 2008. Pp. 71–87.
21. Ontology Engineering with Ontology Design Patterns: Foundations and Applications. Studies on the Semantic Web / Hitzler P., Gangemi A., Janowicz K., Krisnadhi A., Presutti V. (eds.) IOS Press/AKA. 2016.