УДК 004.822

О ВОЗМОЖНОСТЯХ ИСПОЛЬЗОВАНИЯ КОММУНИКАТИВНЫХ ГРАММАТИК И LSPL-ШАБЛОНОВ ДЛЯ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИЙ

© 2015 С.В. Романов, А.А. Сытник, Т.Э. Шульга

Саратовский государственный технический университет им. Ю.А. Гагарина

Поступила в редакцию 30.07.2015

Статья посвящена проблеме автоматического построения онтологий для коллекций текстовых документов. Приведен обзор работ в области автоматического построения онтологий на основе лексико-синтаксических шаблонов (LSPL), выявлены недостатки этого подхода применительно к русскоязычным текстам. Предлагается два варианта совместного использования LSPL-шаблонов и коммуникативных грамматик для решения задачи автоматического построения онтологий.
Ключевые слова: модель представления знаний, онтология, методы автоматического построения онтологий, лексико-синтаксические шаблоны, коммуникативные грамматики.

Конец 20-го века в информатике характеризуется появлением новых подходов к управлению знаниями. При этом одной из самых популярных моделей представления знаний является онтология [1]. В соответствии с определением консорциума W3C под онтологией понимается формальная модель представления знаний в некоторой предметной области. Онтология описывает типы объектов (классы), взаимосвязи между ними (свойства), и способы совместного использования классов и свойств (аксиомы) [2]. Проблемам онтологического описания различных типов ресурсов посвящено большое количество научных работ, например [3, 4, 5, 6].

Построение онтологии для заданной коллекции текстовых документов в конкретной предметной области знаний (например, сборники приказов, архивы историй болезней) является трудоёмкой задачей, которая в основном выполняется экспертами в данной предметной области. Учитывая экспоненциальные темпы роста текстовой информации, представленной в электронном виде, остро встаёт вопрос об облегчении задач экспертов, а именно автоматизации построения начальных версий онтологий [7, 8, 9, 10, 11].

В данной статье рассматривается один из популярных подходов к автоматическому построению онтологий для коллекций англоязычных документов, а именно подход, основанный на использовании лексико-синтаксических шаблонов. Однако, русскоязычные тексты имеют ряд характерных особенностей, такие как падежи имен существительных, обратный порядок слов и т.п., что делает использование данного подхода неэффективным в первоначальном виде. Группа разработчиков создала лексико-синтаксические шаблоны для русского языка [11], но возникает вопрос о точности построения онтологий с помощью данных шаблонов. Для увеличения точности автоматически построенных онтологий авторами предлагается два варианта совместного использования лексико-синтаксических шаблонов и такого средства представления семантики текста как коммуникативные грамматики [12].

Прежде всего, приведем одно из формальных определений онтологии[1].

Онтология - это упорядоченная тройка вида:

$$O = < T, R, F >,$$

где *T — конечное множество терминов (концептов,* понятий, классов) предметной области, которую представляет онтология *O;*

R — конечное множество отношений между понятиями заданной предметной области;

F — конечное множество функций интерпретации (аксиоматизация), заданных на концептах и/или отношениях онтологии *O.*

Таким образом, построение онтологии заключается в выделении концептов (понятий, терминов), отношений между ними и функций их интерпретации.

Простейшими примерами онтологий является простой словарь, где множества R, F — *пусты или простая таксономия* O = <X,{is_a},{}>, у которой X — *множество интерпретируемых терминов, а is_a — отношение «является элементом класса»*

Существуют несколько основных подходов к автоматическому построению онтологий:

- подход на основе лексико-синтаксических шаблонов [7];
- подход на основе системы продукций [8];
- подход на основе статистических методов [9, 10].

В данной статье рассмотрим подход, основанный на использовании LSPL-шаблонов.

Лексико-синтаксические шаблоны (lexico-sytactic patterns) — это лингвистические кон-

Романов Сергей Викторович, аспирант. E-mail: mrtrust@mail.ru
Сытник Александр Александрович, доктор технических наук, профессор, действительный член РАЕН, первый проректор. E-mail: sytnika@sstu.ru
Шульга Татьяна Эриковна, доктор физико-математических наук, доцент. E-mail: shulga@sstu.ru

струкции или схемы, которые показывают семантические связи между терминами и могут быть применены для идентификации формализованных понятий и концептуальных отношений в тексте на естественном языке.

LSPL-шаблоны записываются на языке LSPL (LexicoSyntactic Pattern Language) [13]. Основным элементом LSPL-шаблона является элемент-слово, соответствующий отдельному слову текста и описывающий конкретную словоформу, латинской буквой указывается часть речи слова (например, N – существительное, P – местоимение, V- глагол), а в угловых скобках записываются лексема и указываются значения грамматических признаков, соответствующих данной части речи.

Заметим, что части речи в LSPL-шаблонах для русского языка такие же, как и шаблонах для английского, отсутствуют только артикли (A).

Например, LSPL-шаблон «NP (,NP)* (,)? (и|или) (другие) NP» описывает предложение «Сом, карась, и другие рыбы…». На основе данного соответствия можно сделать вывод, что слова «сом» и «карась» являются терминами-гипонимами по отношению к слову «рыба», где гипоним – это понятие, выражающее частную сущность по отношению к другому, более общему понятию, а следовательно с точки зрения онтологии «сом» и «карась» могут быть рассматриваться как подклассы класса «рыба».

Приведем результаты нескольких научных работ, подтверждающие эффективность использования LSPL-шаблонов для автоматического построения онтологий.

В статье [14] авторы приводят оценку эффективности LSPL-методов для построения онтологий медицинских документов и пополнения существующих онтологий. Были обработаны документы ренгенологии в размере 852,764 шт. для пополнения новой информацией проекта RadLex, который в настоящее время содержит более чем 11000 концептов, а также документы о раковых патологиях в размере 209,997 шт. для пополнения тезауруса Национального Института рака (National Cancer Institute Thesaurus), который в настоящее время содержит более чем 75000 концептов.

На первом этапе работы исследователи из всего объёма документов получили все предложения, соответствующие определенным LSPL-шаблонам. На втором этапе случайно выбранные предложения были направлены экспертам в соответствующих предметных областях, которые оценивали медицинский смысл каждого термина, а также кураторам RadLex и NCIT, которые оценивали термины, с точки зрения возможности их добавления в соответствующие онтологии.

Чтобы оценить полезность LSPL-метода были использованы две оценивающих метрики: suggestion rate (SR) и acceptance rate (AR).

Первая метрика SR показывает процент от полученных терминов, которые являются кандидатами на включение в существующую онтологию. Вторая метрика AR показывает процент терминов, которые могут быть включены в существующую онтологию как новые. Для данных метрик были получены следующие значения: для NCIT метрика SR для концептов – 24%, метрика SR для отношений – 65%, метрика AR для концептов – 21% и метрика AR для отношений – 14%; для RadLex метрика SR для концептов – 37%, метрика SR для отношений – 55%, метрика AR для концептов – 11% и метрика AR для отношений – 44%. Таким образом, было показано, что данный метод обработки текста для пополнения онтологии является эффективным в медицинской области.

В статье [15] авторы описывают процесс создания онтологии с использованием LSPL-шаблонов на основе 733 статей (и соответственно 161585 предложений) из англоязычной WikiPedia. После проверки экспертами была выявлена 77% точность определения концептов и отношений, что также подтверждает эффективность использования LSPL-шаблонов для автоматического построения онтологии.

В статье [5] для генерации онтологии авторы используют LSPL-шаблоны Хейст (Hearst) [16] и DP [17] шаблоны. В работе представлены два приложения: SPRAT и SARDINE, которые используют представленные LSPL-шаблоны для генерации и/или пополнения онтологий. Приложение SARDINE определяет в тексте виды рыб, а приложение SPRAT не имеет привязки к какой-либо предметной области. Для 25 случайно выбранных статей о животных из англоязычной WikiPedia приложение SPRAT сгенерировало 1026 классов, 83,6% из которых были правильными с точки зрения экспертов, 659 подклассов с точностью 76,6%, 23 экземпляра классов с точностью 52,2 % и 55 свойств с точностью 54,5%. Авторы статьи указывают, что примерно такие же результаты давала программа SARDINE.

Таким образом, LSPL-шаблоны позволяют эффективно генерировать онтологии для коллекций англоязычных документов. Однако существующие LSPL-шаблоны не определяют категориальный смысл слов, что является существенным недостатком с точки зрения построения точных онтологий для коллекций русскоязычных документов.

Например, LSPL-шаблон «NP подобно NP» находит гипонимы в текстах. В предложении «Ложка подобно вилке должна лежать рядом с тарелкой» можно выявить что «ложка» и «вилка» являются гипонимами в контексте столовых приборов. Но для предложения «его глаза подобно солнечному свету мелькнули перед ней» данный шаблон неверно определит, что «глаза» и «солнечный свет» являются гипонимами. Средством для на-

хождения семантических связей и определения категориального типа части речи стала коммуникативная грамматика[12].

КОММУНИКАТИВНАЯ ГРАММАТИКА

Коммуникативная грамматика – описание языка, которое раскрывает правила функционирования единиц языка в речи в зависимости от содержания высказывания.

Коммуникативная грамматика призвана быть инструментом для анализа смысла речи, где синтаксические правила используется для определения общего смысла синтаксических конструкций, таких как словосочетание или предложение. Основой коммуникативной грамматики являются синтаксемы. Синтаксемой называется минимальная синтактико-семантическая единица языка, несущая обобщенный категориальный смысл и характеризующаяся взаимодействием морфологических, семантических и функциональных признаков. Синтаксемы присутствует во всех языках и используется для построения более сложных конструкций: предложений, словосочетаний и т.п.

Приведем примеры *значений* синтаксем.

• Аблатив – исходная точка движения (*выйти из комнаты*).

• Агенс – производитель действия (*закон подписан президентом*)

• Адресат – лицо или реже предмет, к которому обращено информативное, донативное или эмотивное действие (*обратиться к президенту*).

• Дестинатив – назначение действия или предмета (*выступать в защиту животных; поехать на лечение*).

• Медиатив – орудие, соединяющее дистантно расположенных участников ситуации или служащее для передачи средства или движения между ними.

Рассмотрим пример предложения: «Митрофанушка не знал, что говорит прозой». Синтаксемами имен существительных в данном предложении являются слово «Митрофанушка» (личное существительное именительного падежа, принадлежащее к классу имен собственных и играющее в данном предложении роль субъекта) и слово «прозой» (существительное в творительном падеже, принадлежащее к классу признаковые и играющее в данном предложении роль медиатива).

Понимание семантических связей между синтаксемами предложения предоставляет возможность более глубокого анализа документа, чем при использовании LSPL-шаблонов. После определения падежа и части речи в данном случае можно с определённой вероятностью говорить о семантических связях между синтаксемами в предложении. Например, «Брат бьёт Машу – Машу бьёт дрожь…Похожие морфологические пары по горизонтали различны по семантико-синтаксической структуре, по составу компонентов…» [12]. В данном случае требуется различать значение субъекта действия (брат) и значения субъекта состояния (дрожь).

Пример показывает, что помимо определения части речи, падежа и других характеристик, требуется определить значение субъекта. Соответственно на передний план выходит задача классификации субъекта, с которой справляется аппарат коммуникативной грамматики.

LSPL-ШАБЛОНЫ И КОММУНИКАТИВНЫЕ ГРАММАТИКИ

LSPL-шаблоны являются эффективный средством для нахождений фраз, словосочетаний по заданным характеристикам в самих шаблонах. То есть, при анализе текстов LSPL-шаблонами можно с высокой точностью находить концепты, связи и отношения между ними, что означает перспективность использования их для обработки коллекций текстовых документов.

Аппарат коммуникативных грамматик позволяет производить семантический анализ фраз, словосочетаний в предложении. При анализе документов имеет место определение не только связей между словами, но и категориального смысла субъектов, что может повысить точность определения концептов и отношения между ними в генерируемой онтологии.

Возникает вопрос о возможности использования LSPL-шаблонов совместно с коммуникативными грамматиками.

Авторами предлагается два следующих подхода к совместному использованию LSPL-шаблонов и коммуникативных грамматик для решения задачи автоматического построения онтологий.

1) Использование аппарата коммуникативных грамматик для верификации концептов онтологии и отношений между ними, полученных в результате анализа текста с помощью LSPL-шаблонов.

2) Создание новых LSPL-шаблонов на основе аппарата коммуникативных грамматик.

Приведем примеры применения данных походов.

ВЕРИФИКАЦИЯ РЕЗУЛЬТАТОВ, ПОЛУЧЕННЫХ С ПОМОЩЬЮ LSPL-ШАБЛОНОВ

Рассмотрим LSPL-шаблон «NP подобно NP», который находит в тексте пары существительных (местоимений), которые являются гипонимами. Пусть дано предложение: «Онегин подобно ветру мчался к Татьяне». В данном случае LSPL-шаблон

определит, что «Онегин» и «ветер» — концепты-гипонимы. Допустим, что уже известно, что «Онегин» является экземпляром класса «Человек», тогда и «ветер» определится как экземпляр класса «Человек», что, очевидно, является неверным. Таким образом, в данном случае будет некорректно определено отношение между концептами онтологии.

На следующем этапе, для проверки правильности выделенных концептов, воспользуемся аппаратом коммуникативных грамматик, которые позволяют определить, что «ветер» — существительное неодушевлённое является субъектом состояния. На основании этого становится ясно, что «Онегин» и «ветер» имеют некоторую общую характеристику в конкретный момент времени, но не общую природу, в отличии, например, от синтаксем «рис» и «гречка» в предложении «Рис подобно гречке обычно варят…».

В итоге, после определения категориального смысла синтаксем можно сказать, что «Онегин» и «ветер» не являются гипонимами.

СОЗДАНИЕ НОВЫХ LSPL-ШАБЛОНОВ

Возникает вопрос о возможности усовершенствования уже существующих LSPL-шаблонов. Из нашего примера можно сделать вывод, что категориальная принадлежность концепта является важной составляющей при анализе текстов. Добавим новую характеристику «cat» – категория концепта для LSPL-шаблонов, которая позволит точнее определять тип связи между концептами.. Для рассмотренного примера получим два новых LSPL-шаблона: «NP1 подобно NP2<NP1.cat=NP2.cat>» и «NP1 подобно NP2<NP1.cat≠NP2.cat>». Первый шаблон будет выявлять связь типа «гипонимы» между концептами, а второй связь типа «метафора».

Таким образом, с помощью коммуникативных грамматик можно производить верификацию концептов, отношений и экземпляров онтологий, полученных с помощью существующих LSPL-шаблонов, а также можно улучшить сами LSPL-шаблоны, за счет введения в них такого компонента как категориальный смысл синтаксем. Такое совместное использование LSPL-шаблонов и аппарата коммуникативных грамматик позволит автоматически генерировать онтологии с высокой точностью.

СПИСОК ЛИТЕРАТУРЫ

1. *Гаврилова Т.А., Хорошевский Ф.В.* Базы знаний интеллектуальных систем. СПб.:Питер, 2000. 382 с.
2. Linked Data Glossary. W3C Working Group Note 27 June 2013 [Электронный ресурс]. URL: http://www.w3.org/TR/2013/NOTE-ld-glossary-20130627/#ontology (дата обращения 04.06.2015)
3. *Осипов Г. С., Тихомиров И. А., Смирнов И. В.* Семантический поиск в сети Интернет средствами поисковой машины Exactus // Труды одиннадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2008: ЛЕНАНД – М., 2008 С. 323-328
4. *Городецкий В.И., Тушканова О.Н.* Онтологии и персонификация профиля пользователя в рекомендующих системах третьего поколения // Онтология проектирования. 2014. № 3 (13). С. 7-31.
5. *Сытник А.А., Вагарина Н.С., Мельникова Н.И.* Онтологическое описание мультимедийных ресурсов в контексте технологий семантического веб // Вестник Саратовского государственного технического университета. 2011. Т. 4. № 2с. С. 202-207.
6. *Валиев М.А., Шульга Т.Э.* Обзор программных продуктов для семантического представления видеоматериалов // Проблемы управления в социально-экономических и технических системах: Сборник научных статей по материалам X Всероссийской научной конференции. Саратов, 2014. С. 3-6.
7. *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.
8. *Найханова Л.В.* Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования: Монография. Улан-Удэ: Изд-во БНЦ СО РАН, 2008. 244 с.
9. Проблема извлечения знаний в информационных системах / *К.А. Амурский, В.В. Дрождин, Ю.Н. Слесарев* // Известия ПГПУ им. В.Г.Белинского. 2010. №18 (22). С. 96-98
10. *Мозжерина Е.С.* Автоматическое построение онтологии по коллекции текстовых документов // Электронные библиотеки: Перспективные Методы и Технологии, Электронные коллекции – RCDL 2011. Воронеж, 2011. С. 293-298.
11. *Большакова Е.И., Васильева Н.Э., Морозов С.С.* Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. М.: Физматлит, 2006. Т. 2. С.506-524
12. *Золотова Г.А., Онипенко Н.К., Сидорова М.Ю.* Коммуникативная грамматика русского языка. М., 2004. 544 с.
13. Lexico-Syntactic Pattern Language. Описание языка. URL: http://lspl.ru/ (дата обращения 14.02.2015).
14. Effectiveness of lexico-syntactic pattern matching for ontology enrichment with clinical documents methods / *K. Liu, W.W. Chapman, G. Savova, C.G. Chute, N. Sioutos, R.S. Crowley* // Inf Med 2010; 49: 397–407, 2011
15. *Klaussner C., Zhekova D.* Lexico-Syntactic Patterns for Automatic Ontology Building, Proceedings of the Student Research Workshop associated with RANLP 2011, pages 109–114, Hissar, Bulgaria, 2011
16. *Maynard D., Funk A., Peters W.* Using Lexico-Syntactic Ontology Design Patterns for ontology creation and population, Proceedings of WOP2009 collocated with ISWC2009, 516, CEUR-WS.org, 2009
17. *Hearst M.A.* Automatic acquisition of hyponyms from

large text corpora. In: Con-ference on Computational Linguistics (COLING'92), Nantes, France, Association for Computational Linguistics, 1992
18. Natural language-based approach for helping in the reuse of ontology design patterns / *G.A. de Cea, A. Gomez-Perez, E.M. Ponsoda, M.C. Suarez-Figueroa* // Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge.

ABOUT USING COMMUNICATIVE GRAMMAR AND LEXICO-SYNTACTIC PATTERNS FOR AUTOMATIC ONTOLOGY BUILDING

© 2015 S.V. Romanov, A.A.Sytnik, T.E. Shulga

Saratov State Technical University named after Yuri Gagarin

Reviewed the method of automatic building ontology using lexico-syntactic patterns, made the review of science works where lexico-syntactic patterns were used, reviewed communicative grammar and their apply together with lexico-syntactic patterns.
Keywords: Automatic building ontology, lexico-syntactic patterns, communicative grammar.

Sergey Romanov, Postgraduate Student.
Email: mrtrust@mail.ru
Alexander Sytnik, Doctor of Technics, Professor, Member of the Russian Academy of Natural Sciences, Vice-Rector. E-mail: sytnika@sstu.ru.
Tatyana Shulga, Doctor of Physics and Mathematics, Associate Professor. E-mail: shulga@sstu.ru