

Композиция ГЛОТЕОН: глоссарий + тезаурус + онтология

09, сентябрь 2011

автор: Клюкин В. Э.

УДК 004.822:004.048

Уральский федеральный университет
имени первого Президента России Б. Н. Ельцина, г. Екатеринбург

vt@dpt.ustu.ru

kve2310@gmail.com

Введение

В настоящее время широко обсуждается использование онтологий в Семантической паутине для поиска и других приложений [1, 2]. Так в [3] было предложено расширить существующую онтологическую модель представления знаний о семантике текста путем добавления в нее представления о Тезаурусе. Нам кажется, что в этом вопросе можно пойти дальше, поскольку тезаурус предназначен для структурирования понятий (преимущественно иерархического), но *определения самих понятий не содержит*. Такое определение понятий содержит другой лингвистический элемент – *Глоссарий*. Рассмотрим подробнее.

Глоссарий и построение системы знаний о тексте

Глоссарий разъясняет лексическое значение того или иного слова. Глоссарий – это также список часто используемых выражений. В частности, рекомендован для использования в проекте Википедия. Но глоссарий содержит не только лексическое понимание ЕЯ-терминов, а главное для нас, он содержит описание сущностей (объективной реальности ПО), выражаемых ЕЯ-терминами. Это позволяет перекинуть недостающий в онтологии мостик между сущностью и ее ЕЯ-описанием. Думается, что это, в конечном счете, позволит расширить онтологию и выразить представление знаний о сущностях и их ЕЯ-описаниях с помощью единого синтаксиса и семантики, а значит, прийти к единому механизму логического вывода. Тогда, сочетая Глоссарий и Онтологию, мы сможем создавать более широкий класс интеллектуальных систем, не выделяя из них лишь ЕЯ-системы.

Как видим, глоссарий прекрасно подходит для разъяснения узкоспециализированных терминов в какой-либо области знаний. Он гарантирует однозначное понимание терминов в данной области и, что особенно важно, перекидывает мостик между объективной реальностью и ее ЕЯ-описанием, специфическим для данной предметной области. А как быть с другими областями? Здесь нам на помощь придет использование *тезауруса*, позволяющего выстроить иерархию. Рассмотрим подробнее.

Тезаурус в современной лингвистике – особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, акронимы, псевдонимы, жаргонные термины и т. п.) между лексическими единицами. Таким образом, тезаурусы, особенно в электронном формате, являются одним из эффективных инструментов для описания отдельных предметных областей. Соотнесения слова с другими понятиями может использоваться для наполнения баз знаний систем искусственного интеллекта.

В искусственном интеллекте и системах автоматического поиска тезаурусы строятся обычно по следующей схеме. Для выбранной предметной области составляется словарь терминов и понятий. Затем эти понятия объединяются в *дескрипторные группы*. Назначение дескрипторной группы – ввести *родо-видовые* отношения между понятиями, а также устранить существующую неоднозначность понимания. Во главе дескрипторной группы ставится *дескриптор* – принятое название рассматриваемого понятия. Дескриптор одновременно выполняет роль однозначного идентификатора всей дескрипторной группы. Вот – простой пример дескрипторной группы из предметной области «Автомобильный транспорт» (полнота группы здесь не учитывается):

АВТОМОБИЛЬ

Род: Автомобильный транспорт

Средства передвижения

Вид: Легковой автомобиль

Грузовой автомобиль

Син: Автомашина

Машина

Тачка

Как видим, родо-видовые отношения дескриптора АВТОМОБИЛЬ легко встроит его в любую систему знаний об Автомобильном транспорте, будь то – семантическая сеть, направленный граф, онтология или формальная логическая система ППП (предикаты первого порядка). Сам дескриптор и его родо-видовые понятия попадут в Глоссарий по данной предметной области. Синонимы, акронимы псевдонимы и т. п. в глоссарий не попадут (минимизация налицо!), хотя при обработке текстов они будут использоваться.

Обратите внимание, например, на жаргонный термин «Тачка». В предметной области «Автомобильный транспорт» он минимизирован в составе дескрипторной группы и не попадет в число дескрипторов. Напротив, в предметной области «Строительство» этот термин займет место дескриптора в соответствующей группе.

Для построения *онтологии* наши Глоссарий и Тезаурус необходимо пополнить сведениями из русской грамматики, позволяющей определить локальный смысл предложения и влияния на

него контекстного окружения. Онтологию в чистом виде рассмотрим здесь только схематично, учитывая большой набор публикаций по ней в последнее время [1].

Онтология (в информатике) – это попытка всеобъемлющей и детальной формализации некоторой области знаний с помощью концептуальной схемы. Обычно такая схема состоит из структуры данных, содержащей все релевантные классы объектов, их связи и правила (теоремы, ограничения), принятые в этой области. Онтологии используются в процессе программирования как форма представления знаний о реальном мире или его части. Основные сферы применения — моделирование бизнес-процессов, семантическая паутина (*Semantic Web*), искусственный интеллект.

Несмотря на всеобъемлющий характер приведенного определения онтологии, хотелось бы заметить, что речь идет *не о самой объективной реальности, а об описании ее* средствами естественного языка. Во-первых, этого часто бывает недостаточно (исключается математика, графика, изображения и др.), во-вторых, описание предусмотрено только на национальном языке (русская грамматика, английская грамматика и т. д.). Т. е. сколько языков, столько и описаний объективной реальности. В одной из следующих работ вопрос о соотношении объективной реальности и ее описании в Семантической паутине будет рассмотрен подробнее.

Здесь следует сделать несколько важных замечаний. Во-первых, фрагменты знания, закрепленные за Онтологией и Тезаурусом, сильно перекрываются. Как отмечено в [3], Тезаурус больше закреплен за лексикой, в проекции на семантику, а Онтология – семантика и прагматика, но все же в проекции на язык. Перед предстоящей композицией эти понятия необходимо четко развести. Во-вторых, построение интеллектуальных систем требует создания хорошей формальной модели предметной области. Онтология же, хотя и называется иногда моделью представления знаний, но это лишь с большой натяжкой для некоторых частных случаев. За онтологией закреплено не представление знаний о самой ПО, а лишь о ее лингвистическом описании, чего естественно недостаточно.

С учетом этих замечаний перейдем к новому образованию – композиции ГЛОТЕОН, объединяющей Глоссарий, Тезаурус и Онтологию, плюс «кое-что еще».

Композиция ГЛОТЕОН

Как замечено в [3], при создании проблемно-ориентированных интеллектуальных систем, оперирующих с ЕЯ-текстами, возникает две основных проблемы: (1) *языковая*, описывающая лингвистическую информацию о проекции ЕЯ на рассматриваемую предметную область, и (2) *сущностная*, описывающая модель предметной области как набор сущностей (объективных реальностей) и межсущностных взаимодействий.

В сегодняшнем понимании Тезаурус ориентирован на лексику и больше приспособлен для решения 1-й проблемы. Однако он не охватывает ее целиком, т. к. не содержит описания

грамматики, морфологии, поверхностного и глубинного синтаксиса, словообразования и др. Онтология, ориентированная сейчас на сущностные связи (семантическую сеть), больше приспособлена для решения 2-й проблемы. Однако она тоже не охватывает ее целиком, т. к. семантическая сеть плохо приспособлена для представления многих практически важных предметных областей. Кроме того, в семантических сетях имеются проблемы с машиной логического вывода.

Поэтому предлагается (с учетом рассмотренного выше понятия Глоссария) создать нечто новое, в которое в качестве базовых компонент входили бы Глоссарий, Тезаурус и Онтология (естественно без перекрытия), а также непокрытые ими части из проблем 1 и 2, описанные выше. Композицию всех этих составляющих назовем ГЛОТЕОН, по имени базовых компонент. Если теперь еще обеспечить «однородность» ГЛОТЕОНа по отношению к языку его описания, то это решит большинство проблем.

В качестве такого языка можно предложить Пролог. Основанный на ППП (подсистема Хорна), он обеспечит описание ГЛОТЕОНа с единых позиций, т. е. его однородность. А кроме того, Пролог имеет не превзойденный на сегодня по быстродействию собственный механизм логического вывода (полиномиальный алгоритм низкого порядка). Использование встроенного механизма вывода позволяет сильно упростить создание интеллектуальных web-приложений.

Пусть даны отображения G – глоссарий и T – тезаурус: $G: P \rightarrow U$, $T: U_g \rightarrow V$, $U_g \subset U$, где P – язык Пролог в интегрированном выражении: синтаксис, семантика, прагматика (включая собственный механизм логического вывода); U – множество понятий; U_g – множество понятий заданной предметной области; V – множество родо-видовых отношений на U .

Тогда для всех $p \in P$ имеет смысл «сквозное отображение»

$$p \mapsto u \mapsto v, u \in U, v \in V, \quad (1)$$

заданное композицией $T^* \circ G$ (см. рисунок):

$$(T^* \circ G)(p) = T^*(G(p)), p \in P. \quad (2)$$

Индекс звездочка (*) здесь и далее означает модификацию тезауруса и онтологии в смысле их семантического разведения, указанного выше.

Обозначим композицию (2) как функцию $\Psi(p)$:

$$\Psi(p) = (T^* \circ G)(p) \quad (3)$$

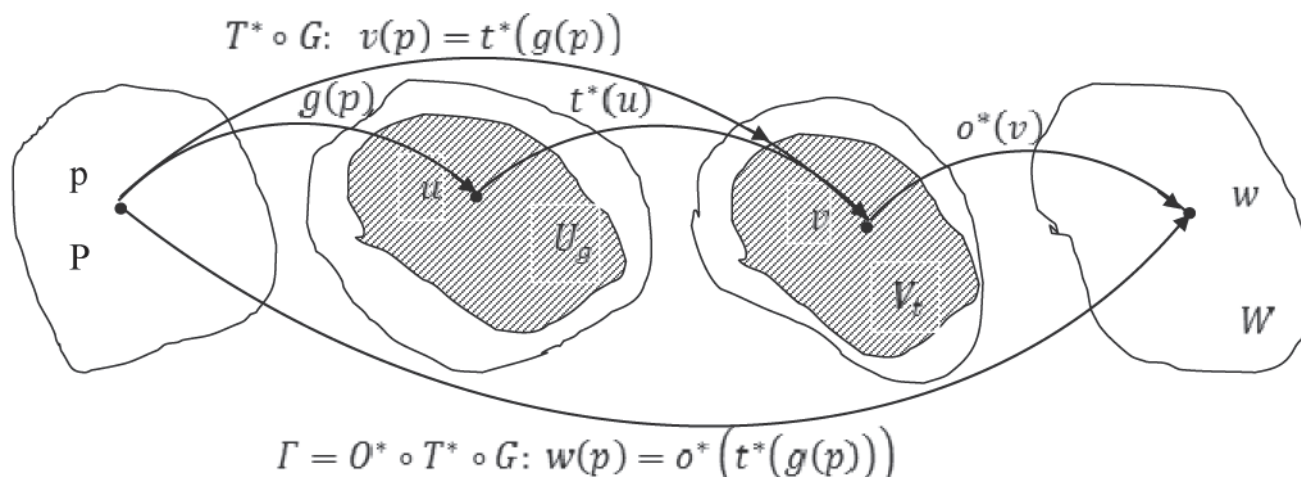
и рассмотрим композицию функций $\Psi \circ O^*$, определяющую сквозное отображение

$$p \mapsto u \mapsto v \mapsto w \text{ из } P \text{ в } W: \quad (4)$$

$$(\Psi \circ O^*)(p) = O^*(\Psi(p)). \quad (5)$$

Подставляя (3) в (5), учитывая (1) и (2), получаем формулу для ГЛОТЕОНа $\Gamma(p)$, описывающего отображение (4):

$$\Gamma(p) = (O^* \circ T^* \circ G)(p) = O^*(T^*(G(p))), p \in P. \quad (6)$$



Полученный результат интерпретируется на рисунке, где $U_g \subset U$, U_g – множество определений сущностей заданной ПО (гlossарий); $V_t \subset V$, V_t – множество родо-видовых отношений заданной ПО (тезаурус); W – множество правил принятия решений, заданное И-ИЛИ – графом, семантической сетью или другой моделью представления знаний (расширенная и разнесенная онтология). Соотношения (4) и (6) показывают однородность и выразимость ГЛОТЕОНа на языке Пролог.

Итак, в составе однородной композиции ГЛОТЕОН зоны ответственности распределены следующим образом:

- Тезаурус отвечает за состав лексики, описывая родо-видовые отношения с помощью дескрипторных групп;
- Онтология описывает представление знаний о ЕЯ-описании ПО, ссылаясь на факты, представленные в Глоссарии и Тезаурусе.
- Глоссарий предназначен для описания (определения) как лексем, так и семантических сущностей.

Синтаксис и семантика Пролога, его выразительная мощь позволяют представить все элементы композиции *однотипно*, а также, построить любые связи взаимодействия ее элементов между собой. Таким образом, именно Пролог образует среду, цементирующую элементы композиции в единое целое – ГЛОТЕОН. И этот ГЛОТЕОН по своей выразительной мощи превосходит онтологию, из которой он зародился. Онтология будет отвечать преимущественно за

отношения между лексемами, представленными в Глоссарии и Тезаурусе, в соответствии с ЕЯ-грамматикой. При этом представлении знаний о грамматике в целом не будет ограничено рамками семантической сети. Пролог позволяет описать направленный граф произвольного вида. Это приведет к расширению возможностей описания грамматических конструкций.

Для целей компьютерной лингвистики, понимания и поиска текста другие составляющие ГЛОТЕОНа практически не будут задействованы. При создании интеллектуальных систем более сложного назначения, могут потребоваться дополнительные знания о модели предметной области. Формализация описания ГЛОТЕОНа с помощью Пролога позволяет справиться и с этой задачей без особых усилий: дополнительные знания можно легко вводить и обрабатывать как и во всякой однородной одноязыковой системе.

Применение ГЛОТЕОНа в Семантической паутине

ГЛОТЕОН прекрасно подходит и для анализа, и для синтеза ЕЯ-текстов, а также их различных гибридов, предназначенных для применения в Семантической паутине. Рассмотрим основные из них.

Анализ при понимании текста и поиске (*текст* → *смысл*). При анализе любая значимая (семантически нагруженная) ЕЯ лексема отражается через ГЛОТЕОН во множество потенциально соответствующих ей смыслов. Для группы связанных лексем их смыслы в общем случае пересекаются, уточняя их возможную общую семантику.

Поскольку у компонентов Онтологии возможность вступления в комбинации ограничена (по образному выражению А. С. Нариньяни «как у атомов при образовании молекул»), то смыслы нескольких текстовых составляющих образуют небольшое число вариантов релевантных семантических структур, которое *сокращается* с расширением контекста соответствующего фрагмента.

Таким образом, до известной степени этот процесс анализа может происходить параллельно, примерно так же, как восприятие зрительного образа. Естественно, в этом случае ГЛОТЕОН должен работать с лингвистическими компонентами, реализующими такие проекции языка, как коммуникативный контекст, структура текста и др., выходящими за традиционные рамки сегодняшнего термина Тезаурус и образующими нечто гораздо более масштабное, включающее ГЛОТЕОН как один из своих компонентов.

Наиболее естественно ГЛОТЕОН вписывается в рамки семантически-ориентированного подхода. При этом процесс от текста к смыслу не связывается с поступательным фронтальным движением "по уровням", как это ранее предполагалось в моделях типа "Текст - Смысл" [3]. Во-первых, в связи с тем, что сами эти стандартизованные уровни в Тезаурусе не присутствуют (напомним, там – только родо/видовые отношения). Во-вторых, потому, что движение от текста к смыслу идет при анализе асинхронно, элементы текста при этом движутся к смыслу самостоятельно, обращаясь

к своему окружению в Тезаурусе или к определению в Глоссарии только в пределах необходимости и затрагивая Онтологию для уточнения своего значения.

Синтез при генерации текста (*смысл* → *текст*). Примерно то же происходит и в процессе синтеза: компоненты смысла ищут свои "ближайшие" текстовые реализации. При этом происходит *расширение* их множества, если этого требует процесс сочетания более элементарных текстовых составляющих в более сложные или же доопределение тех и других в составе создаваемых синтаксических конструкций.

Можно рассмотреть и другие, чаще гибридные, области применения ГЛОТЕОНа:

- *поиск* в Интернете и в наборе текстов конкретных ПО;
- *индексация* материалов ПО;
- *реферирование*;
- *понимание текста* ПО;
- *диалоговые системы* и т. п.

Естественно, что пока еще ГЛОТЕОН не является настолько "совершенными", чтобы одним большим скачком решить все описанные здесь задачи. Необходимо проведение серьезных фрагментарных и глобальных исследований, чтобы пройти этапы от простого к сложному. Планка совершенства ГЛОТЕОНа будет постоянно расти с развитием компьютерной лингвистики, технологии представления и использования знаний.

Стартовые исследования по ГЛОТЕОНу можно начать, например, с его использования при поиске в наборе текстов. Представляется, что этот эксперимент мог бы выглядеть примерно так:

- определить конкретную ПО для разработки web-приложения, например, осуществляющего поиск в наборе текстов;
- создать Глоссарий этой ПО;
- разработать ее Тезаурус;
- создать ее Онтологию;
- обеспечить эффективную взаимосвязь Глоссария, Тезауруса и Онтологии для функционирования ГЛОТЕОНа как единого целого.

Последняя задача, как указывалось, требует согласования двух разных формализмов: компьютерной лингвистики и представления знаний. Думается, что оба эти формализма можно перекрыть одним, более мощным – теорией предикатов первого порядка (ППП). Это позволило бы сразу рассмотреть ГЛОТЕОН как единое целое, с единой «нервной системой». И здесь не надо много изобретать: язык Пролог с его эффективной машиной логического вывода к вашим услугам [4].

Естественно, что направление возбуждения в «нервной системе» ГЛОТЕОНа должно контролироваться текущим процессом. Для того чтобы управлять, надо создать ту среду, которая станет

объектом управления, т.е. такой граф, в котором каждая семантическая связь между его узлами ассоциативно активирована и способна передавать возбуждение в любом нужном направлении.

Прототип такой среды уже удалось создать А. И. Пацкину (Российский НИИ искусственного интеллекта, проект Абриаль) [5]. В эту среду была перенесена ранняя версия Тезауруса Роже. Результаты этого эксперимента кажутся обнадеживающими – это хорошая стартовая точка для последующих исследований.

Итак, в нашей модельной задаче ГЛОТЕОН используется для поиска в наборе текстов. Система получает запрос на ЕЯ с формулировкой темы поиска. Запрос анализируется через ГЛОТЕОН, в результате чего строится его семантический образ. Этот образ активирует семантическое подпространство, покрывающее все возможные семантические составляющие, релевантные в онтологической части ГЛОТЕОНа данному запросу. А семантическое подпространство, в свою очередь, синтезирует на основе глоссальной и тезаурусной частей ГЛОТЕОНа и других лингвистических компонентов все текстовые реализации этих составляющих, которые используются поисковым механизмом в качестве образцов в процессе поиска.

Естественно, что само наличие набора текстов является технологической базой для формирования и Глоссария, и Тезауруса, и Онтологии, не говоря о поисковом механизме. При этом может оказаться, что данная ПО распадется на несколько взаимосвязанных, но содержательно различных составляющих, порождающих соответствующий комплекс вложенных ГЛОТЕОНов, достаточно разнородных, но, возможно, активно взаимодействующих. Однако использование Пролога означает единый синтаксис и единообразные поисковые механизмы (универсальные решатели логических задач), так что интеграция частей этих составляющих в единый интегральный ГЛОТЕОН не должна оказаться слишком сложной.

Впрочем, как замечено в [3], в определенном смысле это и к лучшему. Так декомпозиция онтологии в ГЛОТЕОНе расширяет ее модульность, что правильно и технологически и с точки зрения инженерии знаний. Такая "конфедеративная" организация ГЛОТЕОНа упростит создание систем для тематически разнородных наборов текстов.

Заключение

Глоссарий, содержащий смысловые определения лексем и предметных сущностей, а также перекрывающиеся сегодня понятия Тезауруса и Онтологии функционально и концептуально разведены, а затем из них создана композиция, которая (за счет использования Пролога) представляет собой единое целое, охватывающее как саму модель ПО, так и ее лингвистическое обеспечение. Это новое образование предлагается назвать ГЛОТЕОН в честь исходных составляющих. К сожалению, А. С. Нариньяни уже нет с нами, но идея этого большого ученого жива – родилась композиция ГЛОТЕОН.

Думается, что ГЛОТЕОН, если его предварительные и упрощенные версии появятся в достаточно близком будущем, станет со временем ключевым компонентом распределенных интеллектуальных систем, оперирующих с ЕЯ-текстами.

Благодарности

Автор благодарит А. А. Давыдова, предложившего рассмотреть проблему использования онтологий в интеллектуальных системах, и М. А. Акоева за обсуждение проблемных вопросов использования онтологий.

Литература

1. Лапшин В. А. Онтологии в компьютерных системах. – М.: Научный мир, 2010.
2. Клюкин В. Э. и др. Интеллектуальные агенты и мультиагентные системы для поиска онтологий в задачах принятия решений / Труды XV международной научной конференции молодых учёных ГОУ ВПО УГТУ-УПИ: сборник статей. В 3 ч. Екатеринбург: УГТУ-УПИ, 2009. Ч. 2. С. 278-280.
3. Нариньяни А. С. Кентавр по имени ТЕОН: Тезаурус + Онтология. Российский НИИ искусственного интеллекта, Москва. <http://www.artint.ru/articles/narin/teon.htm>
4. Стерлинг Л., Шапиро Э. Искусство программирования на языке Пролог. – М.: Мир, 1990.
5. Пацкин А. И. Проект Абриаль. <http://www.artint.ru/packin/abrial/index.htm>