OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822

ПОСТРОЕНИЕ МНОГОЯЗЫЧНЫХ ТЕЗАУРУСОВ СРЕДСТВАМИ СЕМАНТИЧЕСКОЙ ТЕХНОЛОГИИ

Загорулько Ю.А.[*, **], Боровикова О.И.[*], Загорулько Г.Б.[*, **]

[*] *Институт систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук, г. Новосибирск, Россия*
[**] *Новосибирский Государственный Университет, г. Новосибирск, Россия*

zagor@iis.nsk.su
olesya@iis.nsk.su
gal@iis.nsk.su

В докладе представлен подход к разработке многоязычного электронного тезауруса для произвольной предметной области, особенностью которого является использование в качестве инструмента разработки формальных и программных средств, предоставляемых семантической технологией, разработанной для построения порталов научных знаний. Благодаря тому, что эта технология базируется на онтологии, обеспечивается не только возможность расширения, целостность и непротиворечивость терминологической системы тезауруса, но и удобный доступ к его контенту.
Ключевые слова: многоязычный тезаурус, концептуальная схема тезауруса, онтология, технология построения порталов научных знаний.

Введение

В связи с глобализацией и интеграцией научного знания часто возникают потребности в разработке многоязычных тезаурусов для различных предметных областей. Однако проблема состоит в том, что на данный момент отсутствуют доступные, гибкие, простые в использовании, но в то же время достаточно мощные средства разработки таких тезаурусов.

При этом под доступными мы понимаем свободно распространяемые или недорогие средства разработки многоязычных тезаурусов, обязательно включающие поддержку разработки русскоязычной версии. Под простыми в использовании – средства разработки, которыми могли бы легко воспользоваться эксперты в предметной области без помощи IT-специалистов и дополнительного обучения. Гибкими мы считаем средства, которые позволяют легко настраивать тезаурус на требуемую предметную область, т.е. в любой момент вводить дополнительные свойства в описания терминов и расширять множество отношений между ними. Важным требованием к таким средствам является поддержание логической целостности терминологической системы тезауруса.

В качестве инструмента построения многоязычного тезауруса, удовлетворяющего всем описанным выше свойствам, в данной работе предлагается использовать методологию и программные компоненты семантической технологии, разработанной для построения порталов научных знаний [Загорулько и др., 2008; Загорулько и др., 2009], которая ранее была применена для создания порталов знаний по археологии [Андреева и др., 2006] и компьютерной лингвистике [Боровикова и др., 2008].

Данная технология базируется на онтологии и предоставляет средства настройки на предметную область и управления контентом информационной системы, а также средства навигации и поиска. Средства настройки на предметную область и поддерживаемая ими методология достаточно хорошо подходят для разработки концептуальной схемы тезауруса, а остальные из перечисленных средств могут выполнять роль его основных программных компонентов, обеспечивающих создание, сопровождение и использование тезауруса.

1. Требования к многоязычному тезаурусу

Перед рассмотрением требований, предъявляемых к многоязычному тезаурусу, определим, какого рода информационный ресурс мы понимаем под тезаурусом.

В современной лингвистике принято следующее определение тезауруса: «Тезаурус (от греч. *сокровище*) — особая разновидность словарей общей или специальной лексики, в которых указаны семантические отношения (синонимы, антонимы, паронимы, гипонимы, гиперонимы и т. п.) между лексическими единицами» [Википедия, 2011]. Из этого определения следует, что главным отличием тезауруса от словарей, в том числе толковых, состоит в том, что в тезаурусе смысл термина представляется главным образом посредством соотнесения его с другими терминами путем установления между ним и этими терминами семантических отношений.

Таким образом, все термины тезауруса оказываются связанными в одну семантическую сеть, представляющую систему знаний о некоторой предметной области (ПрО). Наличие у тезауруса сетевой структуры создает предпосылки для его использования в задачах индексирования и информационного поиска [Лукашевич, 2011].

Для того, чтобы тезаурусом было легко пользоваться и он представлял собой более полную картину моделируемой ПрО, в него включаются определения наиболее важных терминов.

В настоящее время наиболее полно разработаны требования к построению информационно-поисковых тезаурусов (ИПТ), т.е. тезаурусов, ориентированных на индексирование и информационный поиск документов. Эти требования представлены в соответствующих отечественных и международных стандартах [ISO 5964-1985; ГОСТ 7.24-2007; ISO 2788-1986; ГОСТ 7.25-2001; ANSI/NISO Z39.19-2005], которые определяют основные единицы, которые могут включаться в тезаурус, и набор отношений между ними, устанавливают правила сбора массива лексических единиц, формирования словника, построения словарных статей и оформления ИПТ.

По своему составу ИПТ подразделяют на тезаурусы, все единицы которых являются дескрипторами (или предпочтительными терминами), и тезаурусы, включающие как дескрипторы, так и аскрипторы (обычные термины). При этом дескрипторы могут использоваться при индексировании документов и в поисковых запросах, а аскрипторы (как текстовые входы) подлежат замене одним или несколькими дескрипторами.

В зависимости от языковой направленности тезаурусы разделяются на одноязычные и многоязычные. Многоязычный ИПТ содержит термины на нескольких естественных языках и представляет эквивалентные по смыслу понятия на каждом из них.

Согласно стандартам, в словарную статью многоязычного тезауруса входят такие атрибуты, как название, язык термина, релятор (помета, служащая для различения омонимичных терминов), комментарий. Термины тезауруса связываются различными семантическими отношениями, отражающими их место в системе понятий выбранной ПрО. Отношения могут снабжаться набором свойств (в том числе математических), отражающих их семантику в тезаурусе.

Для подтверждения актуальности терминов и ознакомления пользователей тезауруса с практикой их употребления в тезаурус могут также включаться описания источников терминов.

В тезаурусах, разрабатываемых для сложных предметных областей (областей знаний), термины могут соотноситься с подобластями знаний.

2. Концептуальная схема тезауруса

На основе анализа требований, изложенных в разделе 1, была предложена структура тезауруса, основными единицами которого являются термины предметной области, связанные между собой семантическими отношениями, а также описания областей знаний и источников, т.е. текстовых документов или коллекций текстовых документов, в которых термины тезауруса встречаются или определяются. При этом термины подразделяются на дескрипторы и аскрипторы.

Так как предлагаемый в данной работе подход нацелен на создание тезаурусов двойного назначения, т.е. тезаурусов, ориентированных как на решение задач индексирования и информационного поиска, так и на непосредственное использование людьми, желающими обратиться к системе понятий данной предметной области, тезаурусы включают также и определения наиболее важных терминов (дескрипторов).

Как было сказано выше, в качестве инструмента построения многоязычного тезауруса предлагается использовать методологию и программные компоненты семантической технологии, разработанной для построения порталов научных знаний. Так как данная технология базируется на онтологии, то первым шагом построения тезауруса в ее рамках является представление концептуальной схемы тезауруса в виде онтологии, которую мы в дальнейшем будем называть онтологией представления тезауруса. Эта онтология будет не только определять структуры для информационного наполнения (контента) тезауруса, но и служить базисом для организации содержательного доступа к содержащимся в нем знаниям и данным.

Для описания онтологии данная технология предоставляет формализм и поддерживающий его редактор онтологии. С помощью этих средств была построена онтология представления тезауруса O_{Th}, задающая его концептуальную схему:

$$O_{Th} = \langle C, R, T, D, At, P, Axt \rangle,$$

где $C = \{Tr, S_T, S_K\}$ – конечное непустое множество классов, представляющих основные

сущности тезауруса; здесь $Tr = Asc \bigcup Des$ – класс терминов, представляющих понятия ПрО «Компьютерная лингвистика», включающий два подкласса – Asc (термины-аскрипторы) и Des (термины-дескрипторы); S_T – класс источников терминов; S_K – класс областей/подобластей знаний;

$R = R^{TT} \bigcup R^{TST} \bigcup R^{TSK}$ – конечное множество отношений, где

$R^{TT} = \left\{ R_1^{TT}, ..., R_m^{TT} \right\}, R_i^{TT} \subseteq Tr \times Tr$ – конечное множество бинарных отношений, заданных на терминах,

$R^{TST} = \left\{ R^{TSF}, R^{TSP}, R^{TSD} \right\}, R_i^{TST} \subseteq Tr \times S_T$ – бинарные отношения, связывающие термины тезауруса с источниками, причем R^{TSF} связывает термин с источником, где он встречается, R^{TSP} связывает термин с источником, где он встречается в предметном указателе или глоссарии, а R^{TSD} – связывает термин с источником, где дается его определение;

$R^{TSK} = \left\{ R^{SKT}, R^{SKS} \right\}$ – бинарные отношения, служащие для встраивания областей знаний в тезаурус, где $R^{\mathrm{SKT}} \subseteq Tr \times S_K$ связывает термины тезауруса с областями знаний, а $R^{\mathrm{SKS}} \subseteq S_K \times S_K$ – задает иерархию на подобластях знаний;

T – множество стандартных типов;

$D = \{d_1, ..., d_n\}$ – множество доменов $d_i = \{s_1, ..., s_k\}$, где s_i – значение стандартного типа string;

$At = \{at_1, ..., at_w\}$ – конечное множество атрибутов, описывающих свойства основных сущностей тезауруса и отношений между ними; значения этих свойств определены на множестве $T \bigcup D$;

$P = \{P_1, ...P_n\}$ – множество формальных (математических) свойств отношений R^{TT};

Axt – множество аксиом, задающих дополнительные ограничения на связи между терминами.

Таким образом, онтология представления тезауруса описывает классы, представляющие основные сущности тезауруса (термины тезауруса, их источники, области/подобласти знаний), отношения, связывающие объекты этих классов между собой, свойства понятий и отношений, а также аксиомы, определяющие их дополнительную семантику. Кроме того, в онтологии задается множество доменов, т.е. возможных значений атрибутов классов и отношений, что позволяет уменьшить число ошибок при создании/редактировании конкретного тезауруса.

Рассмотрим подробнее классы онтологии, представляющие основные сущности тезауруса.

Класс Tr («Термин») включает следующие атрибуты: *название термина, язык, комментарий, автор статьи*. Заметим, что название термина может задаваться словом, словосочетанием или лексически значимым компонентом сложного слова естественного языка. Автор статьи задается для контроля процесса коллективной разработки тезауруса и в финальную версию тезауруса может не включаться.

Классы Des («Дескриптор») и Asc («Аскриптор»), являясь подклассами класса Tr, наследуют перечисленные выше его атрибуты. Кроме того, класс Des включает следующие дополнительные атрибуты: *релятор, определение термина, признак корневого термина*.

Определение термина поясняет на языке термина его смысл или значение. Заметим, что наличие в тезаурусе определений терминов делает возможным его использование не только в качестве инструмента для ручного или автоматизированного индексирования, но и в качестве источника систематизированных знаний о данной ПрО.

Признак корневого термина указывает на то, что дескриптор находится на самом верхнем уровне одной из иерархии понятий.

Класс S_T («Источник терминов») описывается следующими атрибутами: *название, библиографическая ссылка, язык, тип* (книга, монография, научная статья, документация, учебник, словарь, тезаурус, интернет-ресурс, коллекция текстов и др.), *краткое описание* и *адрес в сети Интернет*. Для коллекции текстов дополнительно может быть задано количество содержащихся в ней текстов и словоупотреблений.

Класс S_K, предназначенный для описания областей/подобластей знаний, включает такие атрибуты как *название* и *описание подобласти*.

В онтологии представлено три типа отношений – отношения между терминами (R^{TT}), отношения между терминами и источниками (R^{TST}) и отношения, служащие для встраивания областей знаний в тезаурус (R^{TSF}).

Главную роль в тезаурусе играют отношения между терминами. Именно они, определяя место каждого термина в системе понятий тезауруса, задают его смысл.

Для того чтобы тезаурус был полезным информационным ресурсом, он должен представлять целостную и непротиворечивую систему понятий ПрО. В единую систему понятия

связываются с помощью семантических отношений, поэтому непротиворечивость системы понятий может быть обеспечена заданием ограничений на связи, устанавливаемые между терминами тезауруса. Такие ограничения могут быть заданы специальными аксиомами, а также путем приписывания отношениям структурных и математических свойств.

Структурные свойства отношений определяют класс их аргументов и тип атрибутов (при их наличии). Для любого отношения из R^{TT} может быть задано одно или несколько непротиворечивых математических свойств из следующего набора: симметричность (*symmetry*), рефлексивность (*reflexivity*), транзитивность (*transitivity*), антирефлексивность (*antireflexivity*), асимметричность (*asymmetry*).

В контексте тезауруса семантика этих свойств описывается следующими аксиомами:

$$\forall x, y \in Tr, x \neq y, R \in R^{TT}, symmetry(R): R(x,y) \rightarrow R(y,x).$$

$$\forall x, y \in Tr, x \neq y, R \in R^{TT}, asymmetry(R): R(x,y) \rightarrow \neg R(y,x).$$

$$\forall x, y, z \in Tr, R \in R^{TT}, transitivity(R): (R(x,y) \,\&\, R(y,z)) \rightarrow R(x,z).$$

$$\forall x, y \in Tr, x \neq y, R \in R^{TT}, reflexivity(R): R(x,y) \rightarrow (R(x,x) \,\&\, R(y,y)).$$

$$\forall x, y \in Tr, x \neq y, R \in R^{TT}, antireflexivity(R): R(x,y) \rightarrow (\neg R(x,x) \,\&\, \neg R(y,y)).$$

Кроме того, для любого отношения из R^{TT} может быть задано обратное (инверсное) отношение. Семантика такого свойства описывается следующей аксиомой:

$$\forall x, y \in Tr, x \neq y, R \in R^{T}, invertibility(R): R, R^{-1} \in R^{T}: R(x,y) \rightarrow R^{-1}(y,x).$$

В соответствии с описанными выше аксиомами: если отношение R обладает свойством симметричности, то для любых терминов x и y наличие связи $R(x,y)$ влечет существование связи $R(y,x)$; если отношение R обладает свойством асимметричности, то при наличии связи $R(x,y)$ между любой парой терминов x и y запрещается существование связи $R(y,x)$; наличие свойства антирефлексивности для $R(x,y)$ накладывает запрет на существование связей $R(x,x)$ и $R(y,y)$; если отношения R_1 и R_2 являются обратными друг к другу, то для любых терминов x и y появление связи $R_1(x,y)$ влечет существование обратной связи $R_2(y,x)$.

Кроме того, для терминов, принадлежащих определенному классу (подмножеству терминов), например, дескрипторам (Des) или аскрипторам (Asc), могут задаваться ограничения на существование или количество каких-либо связей между терминами (x) этого класса и терминами (y,z) этого или других классов терминов. Такие ограничения могут описываться следующими аксиомами:

$$\forall x, y, z \in Des, Des \subseteq Tr, y \neq z, R \in R^{TT}, \text{iniqueness}(R): R(x,y) \rightarrow \neg R(x,z).$$

$$\forall x \in Asc, y_i \in Des, R \in R^{TT}, obligatoriness(R): \exists^{N} R(x,y_i), i = 1,..,N.$$

Первая аксиома задает единственность (*uniqueness*) связи R для некоторого класса терминов (например, Des), т.е. устанавливает, что при существовании у термина x этого класса связи $R(x,y)$ невозможно добавление для него еще одной связи R с другим термином. Вторая аксиома устанавливает обязательность (*obligatoriness*) связи R для некоторого класса терминов; тем самым она контролирует наличие N связей вида $R(x,y)$ для термина x этого класса.

Для отражения семантических связей между понятиями, выражаемыми дескрипторами, устанавливаются иерархические, ассоциативные отношения, а также отношения эквивалентности. Заметим, что если первые два вида отношений устанавливаются между одноязычными дескрипторами, то последнее – между разноязычными.

Между дескрипторами вводятся следующие иерархические отношения:

- Недифференцированная иерархическая связь «Выше», направленная от нижестоящего дескриптора к вышестоящему. Для этого отношения задается обратное отношение «Ниже» и свойство транзитивности.
- Родовидовая связь «ВышеРод», устанавливаемая между двумя дескрипторами, когда объем понятия нижестоящего дескриптора входит в объем понятия вышестоящего дескриптора. Это отношение имеет обратное отношение «НижеВид» и обладает свойством транзитивности.
- «ВышеКласс» служит для задания связи между дескрипторами, представляющими класс понятий и экземпляр этого класса. Данное отношение является нетранзитивным и асимметричным и имеет обратное отношение «ЭкземплярКласса».

• Партонимическая связь «ВышеЦелое», задаваемая между двумя дескрипторами в том случае, когда нижестоящий дескриптор представляет компонент объекта, обозначаемого вышестоящим дескриптором. Для этого отношения задается обратное отношение «НижеЧасть» и свойство транзитивности.

Так как один и тот же дескриптор может одновременно входить в несколько иерархий, построенных не только по различным отношениям, но и по различным основаниям классификации, вводится дополнительный признак «Аспект деления иерархии». Этот признак (в качестве значения атрибута соответствующего отношения) приписывается связям, организующим какое-либо множество дескрипторов в иерархию по одному и тому же отношению и основанию классификации. Тот факт, что какой-либо дескриптор находится в самом верху иерархии, отражается установлением его атрибута *признак корневого термина* в значение *Истина*.

Связи между дескрипторами, отличные от иерархических отношений и отношений синонимии, задаются отношением «Ассоциируется с». Это отношение позволяет задавать произвольные ассоциативные связи между дескрипторами, например, отношения, выражающие зависимости вида «процесс-объект», «причина-следствие» и др. Следует заметить, что если в тезаурусе требуется отразить более богатый набор отношений, специфичных для его ПрО, то такие отношения могут быть введены в онтологию представления тезауруса вместо отношения «Ассоциируется с».

Между одноязычными дескрипторами и аскрипторами устанавливаются отношения синонимии.

Если дескриптор может однозначно во всех контекстах заменить некоторый аскриптор, то он связывается с ним отношением «Синоним». Для этого отношения вводится обратное к нему отношение «Смотри».

Если нет однозначного соответствия между дескрипторами и аскрипторами, то используются следующие отношения:

«Используй альтернативно» – устанавливает связь между аскриптором и множеством альтернативных дескрипторов; обратное ему отношение – «Сравни альтернативный выбор»;

«Используй комбинацию» – позволяет представлять аскриптор комбинацией дескрипторов; обратное ему отношение – «Сравни комбинацию».

Чтобы указать эквивалентность дескрипторов из разных одноязычных версий между ними устанавливается симметричное отношение «Эквивалент на другом языке».

Отношение R^{SKT} («Подобласть знаний») связывает термины тезауруса с областями знаний. Отношение R^{SKS} («Включает») служит для задания иерархий на подобластях знаний.

В онтологии для связи термина с источниками введены три отношения R^{TSF} («Встречается в»), R^{TSP} («Встречается в части документа») и R^{TSD} («Дается определение в»).

Отношение «Встречается в» служит для связывания любого термина с источником; при этом, если источник – коллекция текстов, то в качестве значения специального атрибута этого отношения можно указать частоту встречаемости термина в источнике.

С помощью отношения «Встречается в части документа» можно сообщить, что данный термин встречается в предметном указателе или глоссарии источника, что указывает на важность термина и повышает степень доверия к нему.

С помощью отношения «Дается определение в» термины-дескрипторы, снабженные толкованиями-определениями, связываются с источниками определений.

3. Создание и сопровождение тезауруса

Для задания конкретных терминов, их определений и источников, а также для установления связей между ними используется редактор данных, предоставляемый технологией построения порталов знаний и работающий под управлением онтологии представления тезауруса. Этот редактор обеспечивает экспертов-лингвистов удобным web-интерфейсом для ведения тезауруса. После ввода или редактирования терминов, источников и связей между ними, новая информация становится сразу доступной пользователям тезауруса через пользовательский web-интерфейс.

С целью обеспечения распределенной коллективной разработки тезаурусов используемая технология поддерживает механизм делегирования прав экспертам разных уровней. В соответствии с этим механизмом только эксперты самого высокого уровня могут редактировать структуры тезауруса (с помощью редактора онтологий), а эксперты более низких уровней – только его содержание (с помощью редактора данных).

Кроме того, действует правило, согласно которому редактировать словарную статью может только ее автор. Если кто-то из экспертов захочет внести изменения в «чужую» статью, он должен согласовать свои действия с ее автором.

Следует заметить, что управление контентом тезауруса значительно упрощается благодаря тому, что логическая целостность и непротиворечивость системы понятий тезауруса обеспечивается встроенными в редактор данных специальными механизмами контроля и вывода знаний, работа

которых базируется на аксиомах, описывающих свойства отношений и классов терминов.

В частности, механизмы контроля и вывода знаний контролируют такие свойства отношений, как транзитивность, симметричность, асимметричность, рефлексивность, антирефлексивность, обратимость (наличие обратных отношений), а также ограничения на существование (количество) и обязательность связей. На основе аксиом происходит корректное установление связей между терминами тезауруса, при необходимости осуществляется автоматическое добавление и удаление таких связей.

Например, если рассмотренное выше отношение «Смотри» обладает свойством обратимости и имеет ограничение на число связей (разрешается одна связь), то при связывании аскриптора *x* и дескриптора *y* отношением *Смотри(x,y)* дополнительно будет создана обратная связь *Синоним(y,x)* (если таковой не существовало), а также будет контролироваться запрет на создание связей *Смотри(x,z)* и *Синоним (z,x)* с другим дескриптором *z*.

4. Обеспечение доступа к терминам тезауруса

Доступ к терминам тезауруса и другим его сущностям обеспечивается пользовательским web-интерфейсом, также предоставляемым технологией построения порталов знаний. В этом интерфейсе содержимое тезауруса представляется пользователю в виде сети взаимосвязанных информационных объектов, представляющих термины, описания подобластей знаний, а также описания источников терминов и их определений.

При навигации по тезаурусу обеспечивается возможность выбора необходимых пользователю терминов, детального просмотра их описаний (тезаурусных статей), а также источников (публикаций или коллекций текстов), в которых встречается термин и/или его определение.

Пользователь может указать, какой тип информации его интересует – все термины, дескрипторы, аскрипторы или источники терминов. При этом ему выдается упорядоченный по алфавиту полный список имеющихся в тезаурусе объектов выбранного класса, который отображается в виде html-страницы, содержащей набор ссылок на эти объекты. Дальнейшая навигация по тезаурусу представляет собой процесс перехода от одних элементов тезауруса к другим по заданным между ними связям, отражающим существующие между ними – тезаурусные (между терминами) или библиографические (между терминами и источниками) – отношения.

Например, при просмотре информации о термине-дескрипторе «Семантическое отношение» мы можем видеть значения его атрибутов, а также его связи с другими терминами и источниками терминов (см. Рисунок 1). В частности, мы можем просмотреть определение термина, узнать имя разработчика его тезаурусной статьи. Используя представленные связи термина в качестве элементов навигации, мы можем перейти к просмотру ассоциирующихся с ним терминов, его эквивалента на английском языке, понятий, стоящих выше и ниже его в иерархии, соотнесенным с ним подобластей знаний, а также описаний источников, в которых он встречается и определяется.

Дескриптор	
название	семантическое отношение
язык	русский
релятор	
определение 1	отношение, определяющее смысловые взаимосвязи между наименованиями понятий
автор словарной статьи	Логинова Е.

Связи объекта

Ассоциируется с (RelatedTerm)

Дескриптор
лексическая функция

Встречается дескриптор в (SourceDescriptor)

Источник	частота
Коллекция текстов Диалог 2000-2010	248

Выше (BT)

Дескриптор
семантическое поле

Дается определение в (SourceDef)

Источник	определение
Интернет энциклопедия «Википедия»	1

Ниже (NT)

Дескриптор
антонимия
гиперонимия
гипонимия
конверсивность
омонимия
(Всего: 6)

Подобласть знаний(SubArea)

Подобласть знаний
3.1. Компьютерная лексикология и лексикография - Computational lexicology and lexicography
3.2. Онтологии и тезаурусы - Ontologies and thesauri
CO1.3. Синтаксис и лексическая семантика - Syntax and Lexical semantics

Эквивалент на другом языке (Trans)

Дескриптор
semantic relation

Рисунок 1 – Тезаурусная статья дескриптора «Семантическое отношение»

Для обеспечения доступа к терминам тезауруса из внешних систем разработан программный интерфейс, благодаря которому тезаурус может использоваться при решении задач индексирования и поиска текстовых документов, относящихся к моделируемой тезаурусом предметной области.

Заключение

В докладе представлен подход к разработке многоязычных электронных тезаурусов, общий состав и структура которых удовлетворяют международным и отечественным стандартам. Особенностью подхода является использование в качестве инструмента разработки ранее созданных формальных и программных средств, предоставляемых технологией построения порталов научных знаний. Благодаря тому, что эта технология базируется на онтологии, с помощью которой описывается концептуальная схема данных тезауруса (онтология представления тезауруса),

обеспечивается не только возможность расширения, целостность и непротиворечивость терминологической системы тезауруса, но и удобный доступ к его контенту.

Предлагаемый подход в настоящее время используется для создания русско-английского тезауруса по компьютерной лингвистике [Загорулько и др., 2011; Соколова и др., 2011]. Этот тезаурус разрабатывается как набор одноязычных версий многоязычных ИПТ, при этом выполняется согласованное построение одновременно двух версий тезауруса – русскоязычной и англоязычной.

Заметим, что благодаря наличию средств настройки структуры тезауруса и поддержки ее семантических свойств данный подход может использоваться при построении многоязычных тезаурусов для любых языков и предметных областей.

Работа выполнена при финансовой поддержке РГНФ (проект № 10-04-12108в).

Библиографический список

[Андреева и др., 2006] Андреева О.А., Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Циркин Б.Г., Холюшкин Ю.П. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии // Труды 10-й национальной конференции по искусственному интеллекту с международным участием - КИИ'2006. –Москва: Физматлит, 2006. -Т.3. -С.832-840.

[Боровикова и др., 2008] Боровикова О.И., Загорулько Ю.А., Загорулько Г.Б., Кононенко И.С., Соколова Е.Г. Разработка портала знаний по компьютерной лингвистике // Труды 11-ой национальной конференции по искусственному интеллекту с международным участием КИИ-2008. М.: ЛЕНАНД, 2008. Т.3. С.380-388.

[Википедия, 2011] Википедия. Статья «Тезаурус». http://ru.wikipedia.org/wiki/Тезаурус

[ГОСТ 7.24-2007] ГОСТ 7.24-2007. Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый многоязычный. Состав, структура и основные требования к построению (введен в действие с 1 июля 2008 г.).

[ГОСТ 7.25-2001] ГОСТ 7.25-2001. Система стандартов по информации, библиотечному и издательскому делу. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления (введен в действие с 1 июля 2002 г.)

[Загорулько и др., 2008] Загорулько Ю.А., Боровикова О.И. Подход к построению порталов научных знаний // Автометрия. Новосибирск: 2008. Т. 44. № 1. С. 100–110.

[Загорулько и др., 2009] Загорулько Ю.А. Технология разработки порталов научных знаний // Программные продукты и системы. – 2009. – № 4. –С.25-29.

[Загорулько и др., 2011] Загорулько Ю.А., Боровикова О.И., Кононенко И.С., Соколова Е.Г. Подход к разработке русско-английского тезауруса по компьютерной лингвистике // Труды XIII Всероссийской научной конференции RCDL'2011 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Воронеж, 19-22 октября 2011 г. – Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2011.- С.27-34.

[Лукашевич, 2011] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. –М.: Изд-во МГУ, 2011.

[Соколова и др., 2011] Соколова Е.Г., Семенова С.Ю., Кононенко И.С., Загорулько Ю.А., Кривнова О.Ф., Захаров В.П. Особенности подготовки терминов для русско-английского тезауруса по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог» (Бекасово, 25-29 мая 2011 г.). - М.: РГГУ, 2011, Вып. 10(17), С. 644–655.

[ANSI/NISO Z39.19-2005] ANSI/NISO Z39.19-2005 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (Periodic Review).

[ISO 2788-1986] ISO 2788-1986. Documentation – Guidelines for the establishment and development of monolingual thesauri. Ed. 2.

[ISO 5964-1985] ISO 5964-1985. Documentation - Guidelines for the establishment and development of multilingual thesauri, IDT (Revised by: ISO/DIS 25964-1 Under development).

DEVELOPMENT OF MULTILINGUAL THESAURUS BY MEANS OF SEMANTIC TECHNOLOGY

Zagorulko Yu.A. [*,**], Borovikova O.I. [*,**], Zagorulko G.B. [*]

[*] *A.P. Ershov Institute of Informatics Systems Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

[**] *Novosibirsk State University, Novosibirsk, Russia*

zagor@iis.nsk.su

olesya@iis.nsk.su

gal@iis.nsk.su

The paper presents an approach to the development of multilingual electronic thesaurus for an arbitrary domain. A feature of the approach is the use of the formal and software tools provided by the semantic technology that was developed for the construction of scientific knowledge portals.

Introduction

Because of globalization and integration of scientific knowledge often the need for multilingual thesauri for different subject domains arises. However, the problem lies in the fact that at the moment there are no available, flexible, easy to use, but at the same time sufficiently powerful tools for development of such thesaurus.

As a tool for building a multilingual thesaurus, which satisfies all the above properties, we propose to use the methodology and software components of the semantic technology developed for the building of scientific knowledge portals and which has been previously used to create the knowledge portals on archaeology and computational linguistics.

Main Part

On the basis of the analysis of the requirements of the Russian and international standards that define guidelines and conventions for the construction of information retrieval thesauri, a general structure of the thesaurus and composition of the thesaurus entries was developed. The main units of the thesaurus are the terms of some subject domain, connected by the semantic relations, as well as descriptions of the knowledge areas and sources, i.e. text documents or

collections of text documents which include the terms of the thesaurus or their definitions. The terms of the thesaurus are divided into descriptors and ascriptors (text entries).

The methodology and software components of the semantic technology developed for the building of scientific knowledge portals was used as a tool for building the multilingual thesaurus. As this technology is based on ontology, the first step in the thesaurus building in its framework is a representation of a conceptual scheme of the thesaurus in the form of ontology, which we call the thesaurus representation ontology. This ontology will not only define the structure of information content of the thesaurus, but also serve as a basis for an organization of content-based access to the knowledge and data contained in it.

The developed thesaurus representation ontology describes the classes that represent the basic entities of the thesaurus (thesaurus terms, their sources, the areas/subareas of knowledge), relations linking the objects of these classes, properties of concepts and relations, as well as the axioms that define their additional semantics. In addition, the ontology specifies a set of domains, i.e. the possible values of attributes of the classes and relations to reduce the number of errors when creating/editing the thesaurus.

It should be noted, that the management of the thesaurus content is considerably simplified due to the fact that the logical consistency and integrity of the thesaurus concept system is provided by the special mechanisms of control and inference of knowledge which are built-in the data editor and the work of which is based on axioms describing properties of the relations and classes of terms.

To insert concrete terms, their definitions and sources into the thesaurus, as well as to link them by relations, the data editor provided by the technology of building of knowledge portals and run by an thesaurus representation ontology is used. This editor provides expert-linguists with a convenient web-interface for management of the thesaurus content.

With the aim of providing a distributed collaborative development of the thesaurus the used technology supports the mechanism of delegating the rights to experts of various levels. In accordance with this mechanism only experts of the high level can edit the structure of the thesaurus (using the ontology editor), while the experts of the lower level – only the thesaurus content (using the data editor).

Access to the terms of the thesaurus and its other entities is provided by the user web-interface, which is also provided by the technology of building of the knowledge portals. In this interface, the contents of the thesaurus is presented to the user in the form of a network of interconnected information objects which present the thesaurus terms, descriptions of sub-areas of knowledge, as well as descriptions of the sources of the terms and their definitions.

When navigating the thesaurus a user can a possibility for selection of the required terms, a detailed view of their descriptions (thesaurus entries), as well as sources where the terms and/or them definitions appear.

CONCLUSION

The paper presents an approach to the development of multilingual electronic thesaurus, a general structure of which, composition of the thesaurus entries and set of relations between the terms meet the international and Russian standards. Feature of the approach is the use of formal and program facilities of technology of the building of the scientific knowledge portals as a tool for development. Due to the fact that this technology is based on ontology, with the help of which the conceptual scheme of the thesaurus (the thesaurus representation ontology) is described, the possibility of extension, the integrity and consistency of the terminology system of the thesaurus, and a convenient access to its content are provided.

The proposed approach is currently used for the creation of a Russian-English thesaurus on computational linguistics.

Note that due to the availability of facilities for adjustment of the thesaurus structure and support of its semantic properties this approach can be used for building of a multilingual thesaurus for any language and subject domain.