МЕТОДЫ ФОРМИРОВАИЯ ГЛОССАРИЕВ В УНИВЕРСАЛЬНОМ ТЕРМИНОЛОГИЧЕСКОМ ПРОСТРАНСТВЕ

М. Г. Мальковский, МГУ им. М.В. Ломоносова, malk@cs.msu.su
С. Ю. Соловьев, Межведомственный суперкомпьютерный центр, soloviev@jscc.ru

Ключевые слова: глоссарий, Интернет, понятие, семантика, семантическая сеть, семантическое отношение, термин, универсальное терминологическое пространство

Приводится формальное определение семантической сети научной и деловой лексики, построенной в рамках проекта реализации универсального терминологического пространства. Обсуждаются формальные свойства сети и приводятся результаты ее эмпирического исследования. Описывается метод формирования глоссария по известной вершине сети. Приводится эвристический метод расширения глоссария, позволяющий пополнить начальный глоссарий родственными терминами, более полно раскрывающими его содержание.

1. Универсальное терминологическое пространство

Концептуально универсальное терминологическое пространство (УТП) есть абстрактное хранилище терминологических статей, связанных семантическими отношениями [1,2]. Практически УТП есть семантическая сеть, вершинами которой являются термины научной и деловой лексики, связанные бинарными отношениями (ребрами) типа "это-есть" и "относится-к".

Каждая вершина семантической сети задается:

- термином (строкой символов);

-*возможно*- синонимами термина; и

-*возможно*- определением термина (несколькими строками).

Вершины, в которые входит хотя бы одно ребро, называются понятийными. С понятийными вершинами в УТП связывается дополнительное наименование (наименование понятия; понятие). Обычно это дополнительное наименование является производным от термина понятийной вершины. Например, термину "Кредит" соответствует понятие "Кредиты", термину "Кредитор" – понятие "Кредиторы". По определению в УТП все понятия различны, что позволяет их использовать для однозначной идентификации входящих вершин бинарных отношений.

Таким образом общий вид элементов бинарных отношений семантической сети имеет вид:

это-есть(<Вершина>,<Понятие>)

относится-к(<Вершина>,<Понятие>)

В общем случае термин не может служить для однозначной идентификации произвольных вершин. Так, термин "Андеррайтер" по разному раскрывается в страховании и в биржевом деле.

Андеррайтер - в страховании - лицо, имеющее властные полномочия от руководства страховой компании принимать на страхование предложенные риски ...

Андеррайтер - в биржевом деле - брокер, принявший на себя обязательство разместить ценные бумаги от имени эмитента ...

Вместе с тем, при реализации УТП опытным путем установлено существование стабильной доли уникальных терминов:

0.926 ± 0.004 или $92.6\% \pm 0.4\%$

Независимо от политики формирования УПТ в каждой самостоятельной версии семантической сети 92.6% терминов встречаются ровно один раз, остальные 7.4% терминов могут встречаться 2, 3 и более число раз. Эта закономерность проявилась на всех без исключения версиях семантической сети, начиная с первой, содержащей 2816 вершин-терминов, вплоть до современной версии, содержащей более 30 тысяч терминов.

С целью упрощения формул, примем следующее соглашение:

конструкции

это-есть(<Понятие'>,<Понятие>) и

относится-к(<Понятие'>,<Понятие>)

эквивалентны конструкциям

это-есть(<Вершина'>,<Понятие>) и

относится-к(<Вершина'>,<Понятие>),

где <Вершина'> - понятийная вершина семантической сети, соответствующая наименованию понятия <Понятие'>.

Принятое соглашение позволяет вместо громоздкой конструкции типа

это-есть(статья, соответствующая понятию "Убийства", "Преступления против личности")

использовать запись:

это-есть("Убийства", "Преступления против личности"),

которая содержательно означает, что все многообразие убийств и связанных с ними обстоятельств является собственным подпонятием более широкого понятия "Преступления против личности".

Примеры отношений, представленных в УТП:

относится-к("Кредиторы", "Кредиты")

это-есть(статья "Кредитор по закладной", "Кредиторы")

это-есть("Международные кредиты", "Кредиты")

это-есть(статья "Компенсационный кредит", "Международные кредиты")

2. Методы формирования глоссариев

В проекте www.glossary.ru семантическая сеть используется для генерации глоссариев по запросам пользователей.

Пусть t - некоторая вершина семантической сети. Будем обозначать:

$P(t) = \{ x \mid$ это-есть$(t,x) \}$ - родовые понятия для t;

$S(t) = \{ x \mid$ это-есть$(x,t) \}$ - собственные подпонятия;

$A(t) = \{ x \mid$ относится-к$(x,t) \}$ - свойства t.

Например, для понятийной вершины t = "Средства поверки",

P(t) = { "Средства измерений" },

S(t) = { "Образцовые средства измерений", "Эталоны единиц физических величин",
"Поверочная установка" },

A(t) = { "Погрешность метода поверки" }.

Пусть Y – некоторое множество вершин семантической сети и $F \in \{ P,S, A\}$, будем обозначать F(Y) множество $\{ x \in F(y) \mid y \in Y \}$.

Для построения глоссария G, заданного понятийной вершиной t используется следующая совокупность вершин-статей:

$G(t) = \{ t \} \cup \{ P(t) \cup \{ S(t) \cup \{ A(t)$

Приведенная формула обеспечивает минимум информации о понятии t.

Кроме предъявления глоссария в проекте www.glossary.ru пользователю предоставляется возможность расширить круг статей с помощью механизма наследования свойств [3]. При этом пополнение понимается, как пополнение глоссария терминами, раскрывающими варианты и свойства основного понятия t. В общем случае, в родо-видовых структурах вершины-понятия, обладающие таким свойством, по степени удаленности от вершины t образуют три последовательности:

$П(0) = \{ t \}, \quad П(i) = S(П(i - 1)) \quad (0 < i)$ - подклассы понятия t;

$K(0) = \{ t \}, \quad K(i) = P(П(i - 1)) \quad (0 < i)$ - надклассы для [некоторых] подклассов понятия t;

$C(0) = \{ t \}, \quad C(i) = A(K(i - 1)) \quad (0 < i)$ - наследуемые свойства.

Соответственно последовательность расширений глоссария для понятийной вершины t определяется как:

$E(0) = \{ t \}, \quad E(i) = E(i-1) \cup П(i) \cup K(i) \cup C(i)$

Нетрудно убедится, что $E(1) = G(t)$.

Топология текущей версии семантической сети с указанием терминов и понятий, приписанных вершинам, выложена на сайте www.glossary.ru и доступна в формализованном электронном виде всем исследователям без ограничений.

ЛИТЕРАТУРА

1. Мальковский М.Г., Соловьев С.Ю. Универсальное терминологическое пространство. Труды Международного семинара Диалог'2002 "Компьютерная лингвистика и интеллектуальные технологии", т.1. М:Наука, 2002, с.266-277.
2. Мальковский М.Г., Соловьев С.Ю. Технология формирования универсального терминологического пространства. Сб. "Информационные компьютерные технологии и Интернет в образовании и науке". Москва: изд-во МИИ для инвалидов с нарушением ОДС, 2002, с.54-55.
3. Нильсон Н. Принципы искусственного интеллекта. М:Радио и связь, 1985, 373с.