OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ОТ ТЕРМИНОЛОГИЧЕСКИХ СЕТЕЙ К ТОЛКОВЫМ СЛОВАРЯМ

Мальковский М.Г., Соловьев С.Ю.

Факультет ВМК МГУ имени М.В.Ломоносова, г.Москва, Россия

malk@cs.msu.su

soloviev@glossary.ru

Рассматривается задача преобразования терминологической сети в толковый словарь идеографического типа. Выявляются особенности и трудности поставленной задачи, а также описываются свойства терминологических сетей полезные для ее решения. Обсуждаются принципы организации специального программного обеспечения, поддерживающего интеллектуальную деятельность составителя словаря.
Ключевые слова: терминологическая сеть; толковый словарь; термин; определение.

Введение

Терминологическая сеть [Мальковский и др., 2012] есть продукт деятельности коллектива научных редакторов, занятых систематизацией терминов, представленных своими определениями-толкованиями. С формальной точки зрения терминологическая сеть представляет собой ориентированный граф, узлами которого являются термины, а дугами – экземпляры бинарных отношений из заранее фиксированного набора допустимых отношений. В терминологических сетях:

- каждая дуга $(A, B)_R$ есть упорядоченная пара узлов A и B, помеченная символом отношения R; если для дуги не оговаривается тип отношения, то в записи такой дуги метка опускается;
- набор допустимых отношений обязательно содержит родовидовое отношение P, которому соответствуют дуги $(A, B)_P$, где A – вид, B – род;
- понятийным узлом называется узел, в который заходит хотя бы одна дуга;
- потомками понятийного узла B называются узлы A, связанные с B дугами (A, B);
- предками понятийного узла A называются узлы B, связанные с A дугами (A, B);
- каждый понятийный узел имеет уникальное имя, которое также служит наименованием понятия;
- как правило, наименование понятия есть общее наименование объектов, составляющих его объем: “Музыка”, “Музыкальные инструменты”, “Камерные музыкальные произведения ” и т.д.

Выявление терминов, пригодных для образования понятийных узлов, а также установление отношений между терминами составляет основу интеллектуальной деятельности научных редакторов. В своей работе редакторы используют определения терминов, справочную литературу и собственные знания.

В больших терминологических сетях выделяются так называемые терминологические кластеры, каждый из которых с определенными оговорками можно рассматривать как терминосистему [Шелов, 2003] некоторой проблемной области. Определяющей характеристикой кластера является его центр – выделенный понятийный узел, имя которого отождествляется с наименованием проблемной области. Типичный кластер содержит около 1000 определений терминов, сгруппированных вокруг 150-170 понятий. Количество бинарных отношений кластера достигает 1500 экземпляров, 600 из которых являются родовидовыми.

Переход от терминологической сети заданного кластера к толковому словарю идеографического типа[1] выполняется человеком. Конечная цель составителя словаря состоит в построении последовательности понятийных (тематических) гнезд, каждое из которых – в первом приближении – представляет собой группу взаимосвязанных терминов, раскрывающих через свои определения различные аспекты одного, особо выделенного термина (центрального термина понятийного гнезда; центра гнезда). По формальным признакам

[1] Толковые словари идеографического типа встречаются не часто, однако их уникальная способность служить одновременно и справочником, и учебником выгодно отличает их (см., например, [Годмен, 1988]) от прочей справочной литературы.

понятийное гнездо полностью подобно кластеру, отличие состоит лишь в назначении. Кластер фиксирует понятийный аппарат проблемной области, а гнездо есть часть понятийного аппарата, организованная в виде удобном для чтения. В словаре понятийное гнездо имеет конкретную печатную форму, содержащую наименование, оглавление и термины гнезда, а также схемы и отсылочные списки.

В качестве наименования гнезда, фигурирующим в его заголовке, используется имя центрального термина. Вопросы расположения терминов в гнезде выходят за пределы настоящего изложения, заметим лишь, что здесь не имеет смысла использовать алфавитный порядок, поэтому в интересах навигации все статьи словаря имеют единую нумерацию, и точная ссылка на статью состоит из термина и номера.

Как правило, понятийное гнездо содержит некоторое количество терминов, связанных родовидовыми отношениями с его центром. Традиционно [Гринев-Гриневич, 2008] считается, что родовидовая структура является важнейшим компонентом терминологии. Более того, в развитых толковых словарях[2] родовидовые ссылки являются обязательными атрибутами словарных статей. В относительно компактных гнездах родовидовую иерархию удобно представлять схемой, что позволяет отказаться от атрибутивного метода указания связей между терминами гнезда.

1. Подход к построению словарей

При построении словаря перед его составителем стоят две трудные задачи. Первая задача состоит в необходимости преобразовать терминологическую сеть в структуру словаря. Вторая задача заключается в необходимости переосмыслить значительный объем, вообще говоря, излишне структурированной информации. Подход к разрешению этой проблемной ситуации состоит в создании интерактивного программного инструментария, способного нетривиально поддерживать деятельность составителя толковых словарей. Инструментарий позволяет оперировать так называемыми терминологическими блоками, которые следует рассматривать как прототипы будущих понятийных гнезд. В процессе построения словаря составитель изменяет количество и состав терминологических блоков, предписывая или запрещая построение тех или иных конкретных блоков. Формально терминологический блок представляет собой подсеть, в которой

- выделен один понятийный узел C (центр блока);
- прочие узлы $A_1, \ldots, A_n$ связаны с центром маршрутом (A_i, A_{i1}) (A_{i1}, A_{i2}) … (A_{ik}, C);
- имеется множество узлов, связанных с центром родовидовыми связями (скелет блока);
- потомки каждого понятийного узла либо (i) представлены полностью, либо (ii) полностью отсутствуют.

Понятийные узлы блока, отличные от центра и удовлетворяющие условию (i), называются внутренними узлами. Понятийные узлы блока, удовлетворяющие условию (ii), образуют сечение блока.

В определении терминологического блока зафиксированы желательные свойства понятийного гнезда: во-первых, блок не должен иметь пропусков и, во-вторых, блок должен обладать определенной системой подвидов. Кроме того, последнее свойство терминологических блоков позволяет однозначно доопределять частично заданные блоки минимальным набором узлов. В общем случае, для построения блока достаточно знать его центр и сечение. Базовый алгоритм построения совокупности терминологических блоков по известному терминологическому кластеру имеет следующий вид:

Этап 1. Считать центр кластера центром первого подлежащего построению блока.

Этап 2. Построить скелет блока как совокупность узлов, прямо или косвенно связанных с центром блока родовидовыми связями;

Этап 3. Построить блок (а) доопределением его скелета и (б) включением в его состав потомков тех узлов сечения, которые не имеют собственных подвидов.

Этап 4. Использовать узлы из сечения построенного блока в качестве центров новых терминологических блоков.

Предложенный алгоритм позволяет построить от 50 до 70 начальных терминологических блоков, которые способны сформировать у составителя словаря представления об оптимальных числовых характеристиках потенциальных понятийных гнезд: об общем объеме блока, об объеме его скелета[3] и о количестве раскрытых понятий. Центры блоков-лидеров и блоков-аутсайдеров отмечаются предписывающими и запрещающими метками, которые используются при повторных вычислениях терминологических блоков. Алгоритм повторного построения терминологических блоков отличается от базового алгоритма двумя правилами обработки, имеющими безусловный приоритет.

Правило 1. Если в блок заносится узел, отмеченный предписывающей меткой, то этот узел заносится в сечение, и его потомки в дальнейшем построении блока участия не принимают.

Правило 2. Если в блок заносится узел, отмеченный запрещающей меткой, то потомки этого узла заносятся в блок автоматически.

Описанный процесс манипулирования

[2] См., например, [Михальченко, 2006].

[3] Есть основания предполагать наличие прямой зависимости между общим количеством терминов понятийного гнезда и количеством терминов в его родовидовой структуре.

терминологическими блоками с помощью меток практически неизбежно приводит к повторному использованию в различных блоках одних и тех же узлов. Составитель словаря вынужден противодействовать этому явлению; в противном случае в разных местах словаря появятся точные копии одних и тех же словарных статей. С точки зрения методов противодействия следует различать два случая возникновения узлов-дубликатов.

Случай 1 | Узлы сечений используются в качестве центров иных терминологических блоков. Ситуация штатная; выход из нее состоит в таком редактировании словарной статьи из узла-дубликата, при котором в тексте определения выделяется его анонс. Обработка такого рода известна, она прослеживается, например, в определениях из словаря селевых явлений [Перов, 1996]:

> *Отвал – насыпь из пустых горных пород, некондиционных полезных ископаемых, шлака.* Отвал может размещаться в отрицательных формах рельефа или образовывать положительные. Отвал часто служит очагом зарождения или твердого питания антропогенных селей. (В этом определении анонс выделен курсивом.)

Полное определение используется лишь в одноименном гнезде, анонс же используется во всех остальных случаях. Дополнительно в анонс включается ссылка на гнездо:

> Отвал – насыпь из пустых горных пород, некондиционных полезных ископаемых, шлака. [См. Отвалы]

Случай 2 | В заданном наборе блоков некоторые внутренние и непонятийные узлы одновременно встречаются в двух и более блоках. В дальнейшем такие узлы называются спорными.

2. Подход к обработке спорных узлов

Появление спорных узлов объясняется наличием у них нескольких предков, попадающих в разные блоки. Проблему существования спорных узлов можно разрешить, если для каждого из них указать наиболее предпочтительного предка. Соответствующее решение принимает составитель словаря, а программный инструментарий обеспечивает комфортные условия для принятия адекватного решения. Подобное “разделение обязанностей” характерно для систем приобретения знаний [Осипов, 2013], в частности, для выбора наиболее предпочтительного предка наилучшим образом подходит метод сопоставления.

Программный инструментарий, оставляя окончательное решение на усмотрение составителя словаря, тем не менее, позволяет выполнить формальное исследование контекста того или иного спорного узла с целью выявления его необязательных предков. Исходя из особенностей терминологических сетей [Соловьев, 2008], можно выделить, по крайней мере, пять эвристик для оценки контекстов.

Эвристика 1. Если все потомки B_1, …, B_k понятийного узла A имеют не менее двух предков, то все эти потомки B_1, …, B_k узла A могут считаться необязательными (рисунок 1).

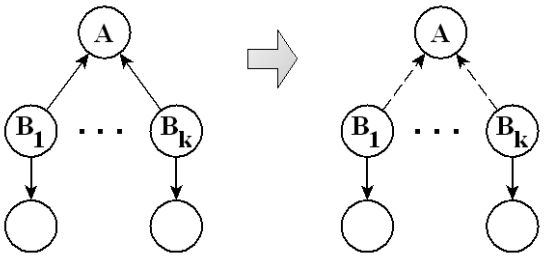

Рисунок 1 – Эвристика 1

Эвристика 2. Если узлы A, B, C связаны отношениями (B, A), (C, A) и (B, C), то потомок B узла C может считаться необязательным (рисунок 2).

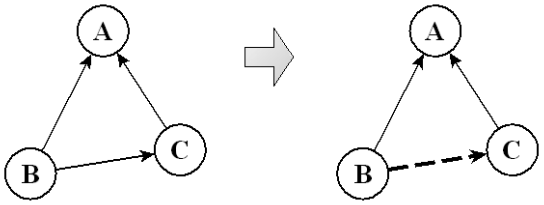

Рисунок 2 – Эвристика 2

Эвристика 3. Если узлы A, B, C, D связаны отношениями (B, A), (C, A)$_P$, (D, B)$_P$ и (D, C), то потомок D узла B может считаться необязательным (рисунок 3).

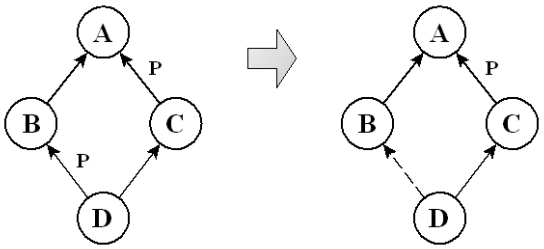

Рисунок 3 – Эвристика 3

Эвристика 4. Если узлы A, B, C, D связаны отношениями (B, A), (C, A), (D, B) и (D, C), то (а) потомок D узла B может считаться необязательным, (б) потомок D узла C может считаться необязательным, и (в) узел D может считаться потомком узла A (рисунок 4).

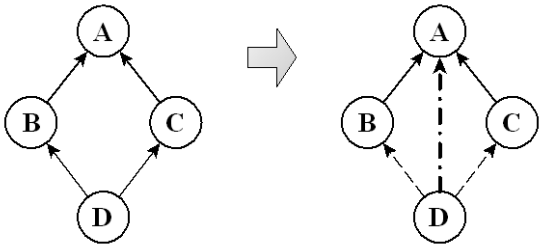

Рисунок 4 – Эвристика 4

Заметим, что эвристики 4 и 3 находятся в

соотношении “правило–исключение”, поэтому эвристика 4 применяется в случае, когда эвристика 3 неприменима.

Эвристика 5. Если среди предков понятийного узла *A* имеется ровно один узел *B*, для которого $h(B) = 1$, то потомок *A* узла *B* может считаться необязательным (рисунок 5). Здесь $h(B)$ – количество понятийных узлов из числа потомков узла *B*.

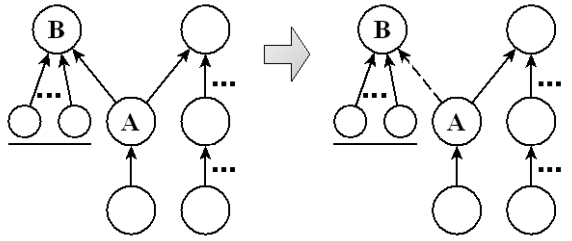

Рисунок 5 – Эвристика 5

При построении терминологических блоков необязательные потомки игнорируются, а в понятийном гнезде имена проигнорированных потомков заносятся в отсылочный список “См. также”, которым завершается печатная форма гнезда.

Заключение

Описанный подход навязывает определенную технологию творческому процессу составления толкового словаря. Ослабить сковывающие рамки технологии можно за счет гибкой реализации соответствующего программного инструментария, скажем, в рамках текстового процессора или системы поддержки творческих картотек. Вместе с тем, программная реализация должна допускать расширение базовых подходов. Например, в толковых словарях изредка встречаются понятийные гнезда с двумя центрами, требующие особых методов обработки. В список вопросов нуждающихся в теоретическом осмыслении, входят: обработка терминов, не имеющих определений, и структурирование терминов в гнезде.

Библиографический список

[Годмен, 1988] Иллюстрированный химический словарь / А.Годмен. – М.: Мир, 1988.

[Гринев-Гриневич, 2008] Введение в терминографию / С.В.Гринев-Гриневич. – М.: ЛИБРОКОМ, 2009.

[Мальковский и др., 2012] Мальковский М.Г., Терминологические сети / М.Г.Мальковский, С.Ю.Соловьев // Материалы конференции OSTIS-2012. – Мн.: БГУИР, 2012. С. 77-82

[Михальченко, 2006] Словарь социолингвистических терминов / Отв. ред. Ю.В.Михальченко. – М.: НИЦ НЯО, 2006.

[Осипов, 2013] Лекции по искусственному интеллекту / Г.С.Осипов. – М.: ЛИБРОКОМ, 2013.

[Перов, 1996] Селевые явления. Терминологический словарь / В.Ф.Перов. – М.: Изд-во Моск. ун-та, 1996.

[Соловьев, 2008] Соловьев С.Ю. Схема и формула глоссария / С.Ю.Соловьев // Труды конференции КИИ-2008, т.2. – М.: ЛЕНАД, 2008. С.157-164

[Шелов, 2003] Термин. Терминологичность. Терминологические определения / С.Д.Шелов. – СПб.: Филологический факультет СПбГУ, 2003.

FROM TERMINOLOGICAL NETWORKS TO THE EXPLANATORY DICTIONARIES

Malkovsky M.G., Soloviev S.Y.

Lomonosov MSU, Moscow, Russia

malk@cs.msu.su

soloviev@glossary.ru

In this article we
– consider the problem of converting existing terminological network into the ideographic explanatory dictionary;
– analyze the characteristics of the data and difficulties in solving the problem;
– describe the properties of terminological network which can be used to solve the problem;
– formulate the principles of software designed to support the dictionary compiler intellectual activity.

Introduction

In our problem the initial data is a cluster of terminological network. Cluster has some useful (for our purposes) numerical and structural characteristics.

Main Part

The compiler creates a dictionary as a set of conceptual slots. Slot describes some subconcept by listing its terms-aspects. Additionally, the slot has a name and contains scheme of relationships between terms. In its activities, the compiler of a dictionary faced with the task of forming slots and with the task of optimizing the number of cross-references between them.

Our idea is to develop a software tool that is able to support the activities of the compiler. The tool allows the compiler to build and adjust terminological blocks. The process ends when the terminological blocks can be converted into conceptual slots. When the compiler checks the blocks he should handle cross-references of two types. References of the first type define connection between the slots, these references require editing the definitions of certain terms. References of the second type are the horizontal connection between the terms, these links affect the contents of terminological blocks. Some heuristics to help handle the second type references.

Conclusion

In essence, the described approach imposes certain technology to the creative process of drawing up an explanatory dictionary. Loosen fetter frame technology can be due to the flexible implementation of appropriate software tools.