ПРОБЛЕМЫ ТЕЗАУРУСА В ИНФОРМАТИКЕ

[1]*А.Н. Сафонова*

Сибирский федеральный университет, 660041 г. Красноярск, пр. Свободный, 79

safonova.nastya1@gmail.com

Доклад посвящен вопросам эффективности применения иерархической структуры понятий, используемых в информатике, обеспечивающей ведение поиска от смысла к лексическим единицам и реконструкции обратного значения. Данные вопросы в своей форме обнаруживают металингвистическую проблематику, поскольку информационный поиск по слову, вынужден исходить из многоконтекстной семантики понятий. Также рассматриваются вопросы значения тезауруса в информатике, проблемы тезауруса на примере информационного поиска, основные проблемы в направлении информационного поиска.

Ключевые слова: тезаурус, информатика, информационный поиск, денотат.

Любая предметная область всегда использует некоторый набор терминов, которые интерпретируются тем или иным понятием или концепцией из этой области. Множество смысловыражающих единиц некоторого языка с заданной на нем системой семантических отношений называется тезаурусом [1]. От греческого тезаурус – это сокровище, сокровищница. Фактически тезаурус определяет семантику некоторого языка, языка конкретной науки или формализованного языка для автоматизированной системы управления. Часто тезаурусы играют важную роль в современных информационных системах и значительно повышают их эффективность.

Различают несколько видов семиотических словарей. Двуязычные словари устанавливают структуру связи знак – знак в разных языках. Толковые словари представляются в виде знак – денотат. Где последний определяется как, объект мысли, отражающий предмет или класс предметов действительности и обозначаемый языковым выражением или языковой единицей – именем [2]. Но в отличии от толковых словарей, тезаурусы имеют обратный вид: денотат – знак.

Впервые термин «тезаурус» в XIII веке использовал Б. Латини в труде «Книга о сокровище», но отцом концепции семантической информации является Ю.А. Шрейдер. Как отмечал Шрейдер – в целом ряде случаев получение информации возможно лишь при наличии определенного запаса или объема предварительной информации по данному вопросу, где предварительная информация вырежется, как

«знание о мире», «некий запас сведений», «представление о мире некоторого наблюдателя» [3].

У каждого языка, предметной области создан свой особенный вид тезауруса. К примеру популярными тезаурусами являются: DUDEN, Roget's, SNOMED, MULTITETS, НАСА, WORDNET [4, 5].

DUDEN – популярный немецкий тезаурус, особенность которого состоит в его конструктивном значении. На каждой его странице представлен некоторый вид человеческой деятельности в картинках. От каждого изображения отходят стрелки с номерами, значения которых описаны на соседней странице по немецкому, русскому и английскому языках.

Также существует наиболее популярный тезаурус Roget's – организован вниз до набора синонимов, что больше всего используется для поиска наиболее подходящего синонима к слову и сопровождается грамматическими сведениями в каждой статье. Выделяют также словарь

SNOMED – большой компьютеризированный тезаурус медицинской терминологии.

Тезаурус НАСА в первую очередь был создан для оптимизации процессов в области аэрокосмических исследований, так как часто проблемой в сложной совместной деятельности является однозначное понимание, то есть обозначение разных ее объектов.

Основные особенности интеллектуального компьютерного тезауруса WORDNET состоят в группировке синонимических слов, называемые синсетами. Группы разбиты на 4 словаря – существительные, прилагательные, глаголы и наречия. Синсеты объединены как в иерархические связи (гипонимы и гиперонимы), так и в отношение антонимии и также меронимии (быть частью чего-л или состоять из частей). Решена также проблема морфологии - слово после обращения к WORDNET возвращается в исходной форме.

Тезаурус применяется во многих областях знаний и пользуется широкой популярностью во всем мире. Более часто тезаурус используется в информатике. Главная необходимость использование тезауруса в информатике – это информационный поиск. Поиск через тезаурус позволит развернуть поисковое слово в необходимых подробностях, опустившись на один уровень ниже в денотатной структуре своего поискового тезауруса, нежели простейший информационный

поиск. Кроме того, с использованием тезауруса запрос может быть автоматически расширен и поиск представится в более полным.

Также следует отметить и другую причину использования тезауруса в информатике – интеграция знаний и повышение эффективности трудовой деятельности за счет оптимизации процесса коммуникации. Все денотаты любого вида деятельности могут быть сведены в понятную пользователю структуру, в которой он легко находит нужный ему денотат, затем его название пользуется им.

Чаще знаки размариваются несколько абстрактно, например – в их отношении между собой, с теми объектами, которые они обозначают, с теми идеями, которые они выражают и то, как они комбинируются друг с другом. Но также следует отметить и то, как знаковые объекты используются в обществе. Одна из проблем тезауруса заключается в передаче информации через знаки и знаковые объекты, которые материализуются в реальном окружающем нас мире, которые воспринимаются по-разному.

Изучение проблем тезаурусов в информатике встречается в настоящее время все чаще, где тезаурус рассматривается с различных точек зрения и получает множество трактовок. Из анализа работ следует, что тезаурус может рассматриваться не только как словарь, но и имеет отличительную особенность, которая четко прослеживается в трудах Ю.А. Шрейдера, где описано различие понятий «абстрактный» тезаурус и «конкретный» тезаурус. Первый в свою очередь представляется как, множество смысловыражающих элементов (слов, словосочетаний и т.д.) некоторого языка с заданными смысловыми отношениями, а второй – задание некоторого абстрактного тезауруса, иначе способ выражения лексических единиц, отношений и самой структуры тезауруса [6].

Также следует представить особенности применения тезауруса в информатике. Одной из задач тезаурусов в данной области является классификация и поиск информационных ресурсов. При этом каждому ресурсу при классификации могут быть сопоставлены одно или более понятий, описываемых терминами в тезаурусе, а пользователь, осуществляющий поиск, может по тезаурусу найти интересующие его понятия в данной предметной области, а также все характеризующие их термины. То есть на основе связей тезауруса происходит расширение поискового запроса (расширение слов запроса синонимичными, более общими или более частными по смыслу терминами). Навигация по связям тезауруса помогает четче сформулировать сам запрос. Различают ряд стандартов разного уровня значимости: для описания

одноязычных тезаурусов и многоязычных. Одноязычные тезаурусы представляют собой набор терминов, связанных между собой соответствующими связями (отношениями), а многоязычные – помимо набора терминов включают связи между эквивалентными терминами на разных языках с такими типамя свизя, как:

- полная эквивалентность;

- неполная эквивалентность – значения терминов не совпадают, но пересекаются;

- частичная эквивалентность – значение одного эквивалента шире, чем значеие другого;

- эквивалентность один ко многим – случай, когда значение одного термина соответствует совокупности значений нескольких терминов.

При наличии в языках-компонентах полностью эквивалентных терминов они считаются представителями одного дескриптора. При отсутствии в языках-компонентах полных эквивалентов для выражения одного и того же понятия в качестве дескриптора в одноязычных версиях используют неполные и частичные эквивалентные дескрипторы. При этом к связям эквивалентности приписывают реляторы или комментарии, описывающие степень эквивалентности.

Существует ряд тезаурусов, основная задача которых не индексация ресурсов, а их классификация. В этом случае основными объектами таких тезаурусов (классификаторов) выступают не термины, а понятия (рубрики), и, часто, идентифицирующие их уникальные идентификаторы (коды классификации). Отношения в таком тезаурусе – не семантические связи между терминами, а характеризующие логику описываемой предметной области отношения между понятиями (рубриками). Примерами таких тезаурусов могут служить тематические классификаторы в разных отраслях науки.

Также важным атрибутом понятия в тезаурусе является комментарий к нему. В тезаурусах-классификаторах, где, по сути, первично понятие, а не термин, комментарий, как правило, также характеризует понятие. Однако, в других тезаурусах комментарий может относиться именно к термину. Например, описывать случаи предпочтительного употребления именно этого синонима перед другими. Таким образом, в разных тезаурусах комментарии могут относиться, как к понятиям, так и к терминам. Выбор зависит от конкретного тезауруса. Универсальная схема данных в информационной системе должна допускать оба варианта применения комментариев.

Различают большое количество тезаурусов, описывающих понятийные и терминологические системы многих предметных областей. Однако, разработка тезауруса для новой предметной области также, как и его пополнение все еще остается большой проблемой. Наиболее качественные тезаурусы создаются вручную, но и существуют различные методы автоматизации (как правило для широких областей знания).

Основными методами являются – методы искусственного интеллекта – извлечение из текста элементов знания (подразумевают большую активность эксперта) [9-10]. Методы делят на:

1 Статистические методы – выделение ключевых слов и словосочетаний.

2 Лингвистические методы – основываются на правилах с использованием шаблонов, на основе которых извлекаются знания из текста (подходы: морфология, синтаксис, семантика).

Главная проблема при автоматическом составлении тезаурусов – большое количество «шума», который следует отсеивать вручную. Что же касается составление тезаурусов «вручную», то здесь проблемы связаны с затратами времени на составление словарей, нехваткой специалистов по автоматизации и экспертов различных областей. Самым эффективным составлением тезаурусных словарей является применение методов в комплексе: использование ручной обработки с автоматическими методами для получения данных большей точности.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Большая советская энциклопедия. http://dic.academic.ru/dic.nsf/bse/.

2 Гуманитарная энциклопедия. Денотат. http://gtmarket.ru/concepts/7039.

3 Семантическая теория по Шрейдеру. http://vikent.ru/enc/1379/.

4 Теория тезауруса. http://ryk-kypc1.narod.ru/tsd.html.

5 Информационные технологии в исскустве. http://pandia.ru/text/78/297/10760-2.php.

6 Шрейдер Ю. А. Тезаурус в информатике и теоретической семантике // Научно-техническая информация.

7 Осокина С.А. Основная лингвистическая теория тезауруса. / Диссертация. Барнаул, 2015.

8 Тезаурус: что это такое. Словарь тезаурус, который больше, чем словарь. http://fb.ru/article/147602/tezaurus-chto-eto-takoe-slovar-tezaurus-kotoryiy-bolshe-chem-slovar;

9 Лукашевич, Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во МГУ, 2011. – 495 с.

10 Гендина, Н. И., Информационно-поисковые тезаурусы: основные виды и области применения // Научные и технические библиотеки. – М.: Государственная публичная научно-техническая библиотека России, 2008 – С. 5-14.

11 М.Х. Нгуен, А.С. Аджиев. Описание и использование тезаурусов в информационных системах, подходы и реализация. / ВЦ РАН Москва, Том 7, выпуск 1, 2004 г., Источник доступа: http://www.elbib.ru/index.phtml/.