Универсальный словарь концептов[1]

Universal dictionary of concepts

Диконов В. Г. (dikonov@iitp.ru)
ИППИ РАН
Богуславский И. М. (bogus@iitp.ru)
UPM/ИППИ РАН
Москва-Мадрид, 2008

В статье представлен универсальный словарь концептов, разрабатываемый в рамках проекта по созданию семантического языка-посредника для глобального обмена информацией. Описываются основные принципы и содержание создаваемого словаря, который может стать общедоступным лингвистически нейтральным ресурсом.

1. Введение

Статья посвящена созданию нового лингвистического ресурса — Универсального словаря концептов (UDC), также именуемого словарём UNL. Он является частью более широкого международного проекта по разработке семантического языка-посредника UNL (Universal Networking Language) [2, 8]. UDC будет использоваться в качестве лексикона этого языка. Хотя словарь тесно связан с языком UNL, он имеет значительную ценность в качестве самостоятельного ресурса и может использоваться для решения различных научных и практических задач, не имеющих отношения к UNL.

2. Универсальный словарь концептов

Основной единицей языка UNL и UDC является концепт — абстрактная семантическая единица, совпадающая со значениями слов, которые выделяются толковыми словарями. Например, согласно данным Merriam-Webster, Collins Cobuild, Oxford и других словарей английского языка слово *baby* может иметь пять значений:

человеческий младенец,
детёныш млекопитающего,
привлекательная девушка,
ребячливый человек,
любимая вещь, идея или проект.

Каждое из них является отдельной лексической единицей UNL и получает уникальный идентификатор — универсальное слово (UW). Обычно для каждого концепта имеется только одно UW.

Словарь не допускает омонимии, то есть ситуации, когда одно UW применялось бы для обозначения разных концептов.

Все концепты заимствуются из естественных языков, а не создаются искусственно. Существование каждого концепта должно быть подкреплено лексикографическими данными какого-нибудь естественного языка или практической необходимостью, например выразить абстрактное грамматическое значение или ввести нетерминальный символ для организации концептов в словаре.

Универсальный словарь концептов стремится включить в себя концептуальные лексиконы всех естественных языков и установить между ними взаимные связи. Если в словаре недостает нужного концепта для описания полисемии слова естественного языка, то следует добавить его, создав новое UW, и определить его связи с другими концептами. Также следует отметить, что каждый концепт имеет свой определённый набор семантических валентностей.

[1] Авторы благодарны РФФИ за частичную финансовую поддержку данной работы (гранты 08-06-00367 и 08-06-00344). В основе данного материала лежит доклад [3] на семинаре проекта MONDILEX (http://www.mondilex.org) «Lexicographic Tools and Techniques» в Москве (на английском языке).

3. Структура словаря

Универсальный словарь концептов должен включать в себя три основных компонента:

1. список концептов, обычно называемый словарём UW;
2. сеть связей между концептами, известная как база знаний UNL (UNLKB)[2];
3. локальные словари, которые связывают концепты со словами различных естественных языков.

3.1. Список концептов

Список концептов включает в себя все имеющиеся в словаре и используемые в языке UNL концепты и существует в виде перечня UW. Различий между UW для полученных из разных языков концептов не проводится. **Все концепты равноправны как отдельные лексические единицы** UNL и включаются в единый список. Вместе с тем, словарь позволит установить источник появления каждого концепта и языки, в которых он имеет прямое лексическое выражение.

Согласно общему принципу каждый концепт должен быть представлен только одним UW. Однако, едва ли возможно полностью избежать ситуаций, когда создаётся несколько разных UW для одного и того же концепта. Такое может происходить по техническим и организационным причинам в децентрализованном сообществе. Словарь должен иметь средства для разрешения подобных коллизий.

В простейшем из случаев уже существующее UW изменяется для того, чтобы исправить ошибку, обеспечить лучшее описание концепта или дополнить UW недостающей информацией. Прежнюю версию UW нельзя удалить немедленно, потому что она может быть использована в существующих UNL-текстах (или на неё могут ссылаться другие лингвистические ресурсы). Простое удаление сделало бы такие тексты несовместимыми со словарём. Словарь должен иметь механизм хранения истории изменения каждого UW, позволяющий отслеживать каждую зарегистрированную версию UW и не допускать повторного введения устаревших UW в словарь.

Ещё одним источником нескольких UW для обозначения одного концепта является сама природа человеческого языка и процессов категоризации. Каждый естественный язык содержит определённое количество полных синонимов, которые могут со временем разойтись в своём значении, например *everyone* и *everybody* в английском языке. Составить их исчерпывающий и точный список чрезвычайно трудно. В результате, на основе таких слов неизбежно будут возникать дополнительные UW для концептов, которые уже имеют своё UW.

Оба процесса создают группы UW, напоминающие синсеты в ресурсах семейства Wordnet [6]. Такие группы следует выделять среди массы синонимов, рассматриваемых как сходные, но разные концепты.

3.2. Семантическая сеть

Концепты образуют семантическую сеть, связанную отношениями гиперонимии, меронимии, конкретизации, синонимии, антонимии, ассоциации и отношений, которые описывают семантические валентности концептов. Назначение семантической сети — предоставить по возможности правильное и объективное описание связей между концептами, которые существуют в человеческом языке и сознании.

Семантическая сеть состоит из трёх различных структур, формируемых а) онтологическими отношениями, которые организуют концепты в группы согласно различным семантическим классификациям, б) семантическими отношениями, которые фиксируют подобие или различие между концептами, и в) аргументными отношениями, которые указывают набор валентностей каждого концепта и возможные классы их заполнителей.

3.2.1. Онтологическая структура

Онтологическая структура состоит из UNL-отношений **icl** (гиперонимия), **pof** (меронимия) и **iof** (конкретизация). Дополнительно могут быть использованы некоторые другие отношения, в частности **val** (значение параметра) и **fld** (область знаний).

Отношения *icl* и *iof* имеют привилегированный статус, так как хотя бы одно из них обязательно присутствует в каждом UW. С их помощью фиксируется принадлежность выражаемого UW концепта к одному или нескольким общим онтологическим классам. Каждый концепт должен быть связан со всеми классами, непосредственным представителем которых он является. Результатом является обладающая свойством иерархичности сеть онтологических отношений, встроенная в сеть из прочих отношений. Гиперонимические классы иерархичны по своей природе. Со значительной степенью упрощения они могут быть выстроены в форме дерева, хотя реальные связи между классами могут быть более сложными (см. рисунок

[2] В более ранних публикациях на связанные с UNL темы UNLKB может именоваться «Master entries dictionary» (словарь полных вариантов). Это название связано с идеей ввести развёрнутые варианты UW (Master Definition), которые бы включали в свой список ограничений все связи концепта с любыми другими концептами. В настоящее время полные варианты UW не используются, но их легко будет получить из UNLKB.

5 ниже). Словарь UDC предлагает более гибкий и реалистичный метод представления отношений между концептами, чем обычное дерево. Возникающая в результате структура оказывается гибридной. В ней совмещаются свойства дерева и сети. Цепи отношений могут разделяться и затем соединяться вновь, как показано на рисунке 1, но вместе с тем структура имеет общую исходную точку или корень.

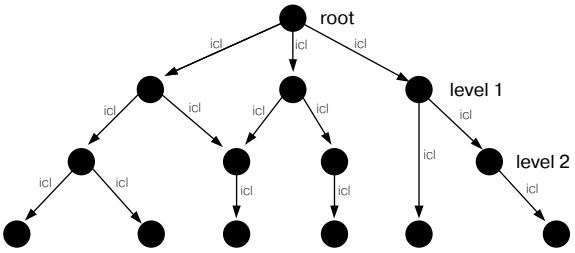

Рис. 1. Онтологическая структура

Абстрактный корневой класс называется «uw» (произвольный концепт). Он подразделяется на более узкие абстрактные классы объектов, свойств, действий, событий и т.п. Можно говорить об уровнях онтологической структуры в рамках одной цепи гиперонимов между концептом и корневым узлом, но концепт может одновременно относиться к нескольким уровням или ветвям структуры.

Онтологические отношения позволяют прослеживать соотношения объёмов понятий и находить обобщающие термины при деконверсии, если в целевом языке нет точного переводного эквивалента Например: при переводе русского слова *жениться*, которое буквально означает *«обрести жену»* и не имеет точного эквивалента в английском, следует заменить соответствующий концепт на более общий *«вступить в брак»*, который имеет прямой перевод на английский язык (см. выше UW с заглавным словом *marry*) .

3.2.2. Семантическая структура

Семантическая структура устроена иначе. Она состоит из семантических отношений **equ** (синонимия), **ant** (антонимия) и **com** (ассоциация). Отношение equ не позволяет различить полные и квази-синонимы. Поэтому его можно дополнить другим выразительным средством, позволяющими маркировать группы UW, которые обозначают один и тот же концепт. Семантические отношения связывают концепты в группы и не образуют никакой иерархии. Возникающая в результате структура, как показано на рисунке 2, является децентрализованной сетью.

В отличие от онтологической структуры, семантическая не обязательно должна быть связной. Она может состоять из многих изолированных фрагментов.

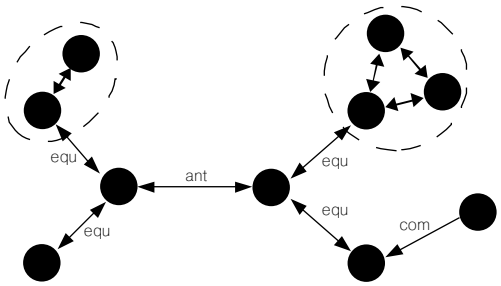

Рис. 2. Фрагмент семантической структуры

3.2.3. Аргументная структура

Аргументная структура является набором аргументных отношений, например **agt** (агент), **obj** (объект), **ptn** (партнер), **ben** (бенефициар), **plt** (место назначения), **src** (источник), **gol** (конечное состояние) и т.п. Эти отношения связывают каждый концепт с концептами абстрактных классов, представители которых обычно заполняют соответствующую отношению валентность. В большинстве случаев аргументные отношения указывают на концепты, которые принадлежат к сравнительно компактной группе наиболее абстрактных онтологических классов, расположенных близко к корню онтологической структуры (рисунок 3).

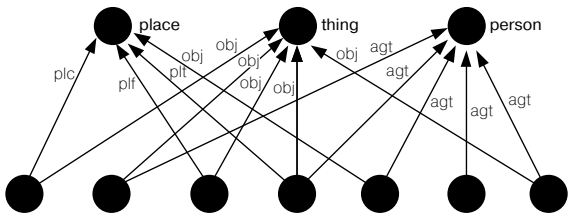

Рис. 3. Аргументная структура

Все три структуры объединяют одни и те же концепты и накладываются друг на друга. Вместе они составляют семантическую сеть UDC.

3.3. Локальные словари

Локальные словари хранят информацию о соответствии концептов и лексики конкретного естественного языка. Для каждого поддерживаемого инфраструктурой UNL языка должен существовать свой локальный словарь. Простейший локальный словарь может быть просто списком пар концептов и их переводов на естественный язык. К словам естественного языка может быть добавлена морфологическая и грамматическая информация, а также любые другие полезные сведения.

Перевод концепта на естественный язык может быть не одним словом, а словосочетанием или целой фразой. Некоторые концепты могут выражаться однословно в одном языке и словосочетаниями или аббревиатурами в другом, например *старшеклассник — senior pupil* или *важная персона — VIP*.

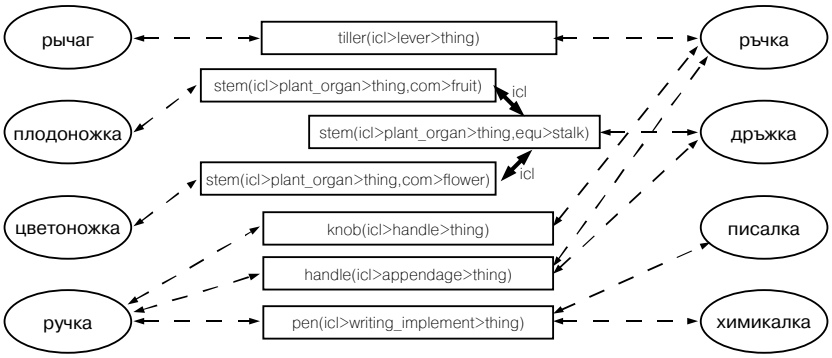

Рис. 4. Концепты и возможные связи нескольких русских и болгарских слов

Однако, перевести некоторые концепты на отдельные языки не удаётся даже описательно. Если необходимо перевести такой концепт, следует при помощи семантической сети найти ближайший более общий термин или наоборот — более узкий. На рисунке 4 представлен подобный пример. На рисунке показаны связи между русскими (слева) словами *ручка*, *рычаг*, *плодоножка*, *цветоножка* и их болгарскими эквивалентами (справа), где UW используются в качестве промежуточного средства представления значений этих слов. В русском языке нет прямого соответствия болгарскому слову *дръжка* в значении *орган растения поддерживающий цветок или плод*. Для нахождения перевода нужно проследить онтологические (icl) связи, которые ведут к концептам *плодоножка* и *цветоножка*. Кроме того, концепт *ручка для письма* имеет два возможных перевода на болгарский.

4. Универсальный словарь концептов и Wordnet

Универсальный словарь концептов похож на ресурсы семейства Wordnet во многих важных аспектах. Оба словаря используют базовое понятие концепта и определяют совпадающие типы связей между своими словарными единицами. Большая часть данных универсального словаря на момент написания статьи была получена из Princeton Wordnet [1]. Ещё больше информации, включая новые концепты и связи между ними [5], можно импортировать из других ресурсов семейства Wordnet. Однако, между UDC и ресурсами Wordnet имеются и важные различия.

4.1. Связь с естественными языками

Каждый Wordnet описывает лексику определённого естественного языка. Разные ресурсы семейства Wordnet могут быть связаны между собой при помощи межъязыковых индексов (ILI). Эти индексы описывают связи между синсетами определённых версий Princeton Wordnet (часто старых версий 1,5 или 1,6) и Wordnet-ресурсами других естественных языков. Однако, ILI-индексы играют вспомогательную роль. Только некоторые из неанглийских Wordnet-ресурсов имеют привязку к Princeton Wordnet. Кроме того, эти связи устаревают с выходом его новых версий.

Универсальный словарь концептов можно сравнивать с несколькими связанными посредством ILI словарями Wordnet, но индексы ILI связывают не все языки со всеми, а только отдельные пары языков, причем, как правило, один язык в такой паре — английский. В универсальном словаре концептов нет явного предпочтения концептуального лексикона одного из естественных языков как исчерпывающего эталона для всех прочих. Вместо этого основное внимание направлено на формирование единого общего набора концептов и установление связей между ними. Связи со словами естественных языков определяются в рамках локальных словарей. Они не будут теряться при изменениях списка концептов и структуры семантической сети.

UW состоят из заглавного слова и набора ограничителей. В качестве заглавных слов, как правило, используются слова английского языка. Однако, это не значит, что словарь использует английский как посредник или эталон при описании других языков. Он был выбран в качестве **основного источника** заглавных слов по сугубо практическим причинам — как единственный язык, знания которого можно потребовать от всех членов коллектива разработчиков. В тоже время, когда для концепта иного языка нет точного эквивалента в английском, с помощью ограничителей можно модифицировать значение английского слова и обеспечить точное описание. Кроме того, не все заглавные слова UW происходят из английского языка.

Концепты происходящие из любых языков могут быть непосредственно связаны друг с другом и служить основой для создания новых UW или ограничителей для описания любых других концептов. Например:

samovar(icl>boiler>concrete_thing,com>tea)
tula_samovar(icl>***samovar***>concrete_thing,com>tula(iof>city))

***sauna**(icl>sweating_room>place,com>finnish,com>dry)*
***parilka**(icl>sweating_room>place,com>russian,com>steam)*
*venik(icl>massage_tool>...com>**parilka(icl>sweating_room)**)*

Если число специфичных для других языков мира концептов будет расти, оснований для утверждения об особой роли английского в UDC будет становиться всё меньше.

4.2. Иерархические структуры

Словари семейства Wordnet организуют именные и глагольные концепты в гиперонимические деревья. Структуры такого рода удобны для поиска и анализа, однако древесная классификация в чистом виде не поддерживает частично пересекающиеся классы. Деревья могут без оговорок и упрощений применяться только для самых верхних уровней полной лингвистической онтологии. Например, в Princeton Wordnet имеются концепты *(теннисной) ракетки* и *(хоккейной) шайбы*, а также класс «*спортивный инвентарь*». При этом *ракетка* является членом класса спортивного инвентаря, а *шайба* — нет. Вместо этого она включена в класс «*дискообразные предметы*». Перемещение концепта *шайба* в класс «*спортивный инвентарь*» в чисто древесной структуре приведёт к потере информации о том, что этот предмет имеет дискообразную форму.

Универсальный словарь концептов стремится реализовать другой менее формально ограниченный подход. Базовая онтологическая структура является сетевым графом, с некоторыми чертами древесности. Наличие у концепта нескольких родительских узлов является допустимым. Это позволяет давать более подробное описание каждого отдельного концепта и реализовывать более полные и детальные классификации множеств концептов. Каждый концепт должен быть связан со всеми возможными непосредственными гиперонимами. Например, слово *суши* в Wordnet непосредственно входит в класс *блюдо* (приготовленная пища). Предположим, что мы хотим ввести дополнительную классификацию блюд по национальности (суши — блюдо японской кухни) и основному ингредиенту (суши делается из рыбы). Определить, какой из двух новых классов выше в иерархии, невозможно, потому что они соответствуют пересекающимся множествам объектов (Рисунок 5)[3].

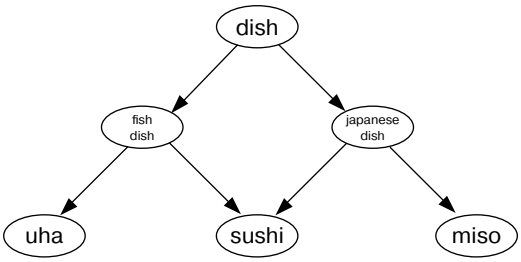

Рис. 5. Несколько родительских классов

Использование сетевой структуры вместо древесной имеет некоторые последствия. Древесная структура позволяет с полной уверенностью проследить цепочку гиперонимов каждого концепта до корня дерева даже при наличии петель, как в Wordnet. Гибридная сетевая структура допускает множество равнозначных цепочек, приводящих к различным и взаимоисключающим классам на высших уровнях иерархии гиперонимов. Это может привести к неопределённости и путанице. Например, класс «*functional thing*», одним из представителей которого является концепт *hammer*, может одновременно входить в классы «*abstract thing*» и «*concrete thing*», тем самым допуская гипотезу, что *молоток* является нематериальным объектом! Для UW эта проблема может сниматься путём добавления дополнительной связи с надлежащим вершинным классом.

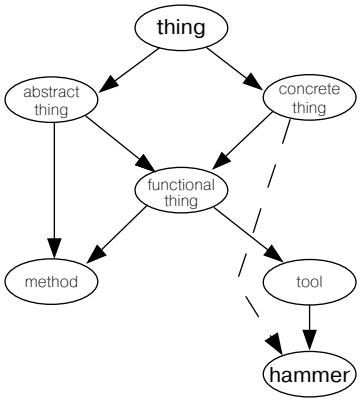

Рис. 6. Дополнительная связь с правильным вершинным классом

Согласно рисунку 6 UW концепта молоток должно выглядеть так: *hammer(icl>tool>**concrete_thing**)*. Когда известны оба конца цепи гиперонимов, становится возможным прослеживать онтологические отношения в гибридной сетевой структуре между любым концептом и соответствующим вершинным классом. Это позволяет получить полную иерархию классов.

4.3. Прочие особенности

В отличие от Wordnet универсальный словарь концептов не ограничивается определёнными частями речи. Он включает в себя полный набор концептов, соответствующих предлогам, союзам и словам со специальными грамматическими функ-

[3] Princeton Wordnet предлагает способ включить синсет одновременно в несколько классов на одном уровне иерархии, но такое расширенное описание не стало повсеместным. Например, слово *key* в значении «*килограмм наркотического вещества*» одновременно включен в находящиеся на одном уровне классы «*единица массы*» и «*единица измерения метрической системы*», а уже на следующем уровне две гиперонимические цепи объединяются в класс «*единицы измерения*».

циями, например, модальным глаголам. Это связано с тем, что в UNL нет явного деления на части речи и каких-либо ограничений круга значений, которые могут быть выражены с помощью UW.

Универсальный словарь концептов предоставляет информацию о семантических валентностях своих единиц. Валентности обозначаются отношениями языка UNL. Для каждой из них указывается наиболее общий онтологический класс, представители которого обычно заполняют эту валентность. В ресурсах семейства Wordnet могут присутствовать сведения о типичном контекстном окружении членов синсета, но единого подхода не существует. Так, Princeton Wordnet описывает модели управления глаголов (sentence frames) без какой-либо семантической классификации связей глагола с актантами, а сами актанты подразделяются только на два класса «somebody» и «something». Однако, в пока ещё неопубликованном словаре Russnet [7], который является наиболее многообещающим аналогом Wordnet для русского языка, описание семантических валентностей ожидается [6].

Некоторые словари семейства Wordnet описывают также синтаксические свойства слов, такие как часть речи, род, одушевлённость и т.п. [9], в то время как другие опираются на программы морфологического анализа и синтеза. В универсальном словаре концептов такой информации нет, так как для универсального семантического языка она не имеет смысла. Соответствующие сведения о словах естественных языков могут быть включены в локальные словари.

5. Развитие словаря

Важно, чтобы процесс развития словаря следовал принципам **разделения труда, постепенности, использования уже накопленных данных и децентрализации**. Поскольку ни один исследователь или коллектив не имеет достаточных ресурсов и знаний для решения задачи в целом, наилучшей формой организации представляется модель открытого сообщества.

Каждый раз, когда накапливается значительный пакет изменений, и нет формальных возражений против них, следует делать очередной срез словаря и публиковать его в качестве новой версии. С этого момента все участники проекта должны обновить собственные производные ресурсы для использования новой версии словаря концептов.

Универсальный словарь концептов будет опубликован под свободной лицензией, как только будет завершена работа над первой версией. Это подразумевает, что данные можно будет распространять и использовать для любых научных и личных целей. Каждый будет иметь право расширять ресурс и исправлять ошибки при условии, что все изменения будут возвращены сообществу пользователей и редакторов словаря. Качество предлагаемых к включению в словарь новых данных должно проверяться экспертами.

Литература

1. *Bekios J., Boguslavsky I., Cardeñosa J., Gallardo C.* An Efficient Method for Building Multilingual Lexical Resources // Proceedings of the Fifth International Conference Information Research and Applications i.TECH 2007, T. 1. Sofia.: 2007. C. 39–45.
2. *Boguslavsky I., Cardeñosa J., Gallardo C., Iraola L.* The UNL Initiative: An Overview // Computational Linguistics and Intelligent Text Processing.2005.
3. *Boguslavsky I. M., Dikonov V. G.* Universal Dictionary of Concepts // Proceedings of the first MONDILEX workshop «Lexicographic Tools and Techniques». M.: 2008. C.31-42.
4. *Fellbaum, C.* WordNet: An Electronic Lexical Database // MIT Press.1998.
5. *Iraola L.* Using WordNet for linking UWs to the UNL UW // International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies.Alexandria.: 2003
6. *Азарова И. В.* Схемы управления в грамматике и рамки валентностей в RussNet // http:// project.phil.pu.ru/RussNet. 2005
7. *Азарова И. В., Митрофанова О. А., Синопальникова А. А.* Компьютерный тезаурус русского языка типа WordNet // Материалы конференции Диалог 2003. М.: 2003.
8. *Веб-страница* проекта UNL http://www.undl.org.
9. *Сухоногов А. М., Яблонский С. А.* Разработка русского WordNet // RCDL2004.Пущино.: 2004.