Вычисление русскоязычного значения английских лексем

Богдановский А.Е. (ab_2003@hotbox.ru)

Санкт-Петербургский Государственный Университет

Введение.

Вычисление русскоязычного эквивалента значения английских лексем является естественным продолжением семантического анализа текстов на русском языке. В данной работе рассматривается главный инструмент формализации структуры данных – классификация объектов семантической модели языка. В системе классификации многомерных понятий большое значение имеет описание межклассовых и внутриклассовых отношений. На основании выявленных отношений конструируются сложные объекты, описывающие целые предложения. Объекты, соответствующие отдельным понятиям, позволяют определить лексемы, описывающие данные понятия. Из полученных лексем возможно построение предложения на другом языке, отличном от языка анализируемого текста. Обсуждаемые проблемы описываются в рамках функциональной модели естественного языка, которая позволяет построить алгоритм вычисления значения слова в контексте целого предложения.

1. Иерархическая система понятий естественного языка

Классификация является одним из методов формализации эмпирических данных. Это метод многоступенчатого деления логического объема понятия или какой-либо совокупности единиц на систему соподчиненных понятий или классов объектов. Цель классификации – определение места в системе любой единицы, и тем самым установление между ними наличия связей [4]. При классификации сложных многомерных объектов происходит их кластеризация на основании одного из критериев (свойство или набор свойств), который выбран в качестве главного для данного кластера. Можно выделить несколько таких критериев для одного объекта, что будет соответствовать различным вариантам разбиения. Тогда уникальность каждого объекта будет определяться набором классов.

Иерархическая модель решает проблему идентификации объектов лишь частично. Она позволяет сопоставить объекту различные классы, на отношения между которыми наложены ограничения. Отношения должны носить родо-видовой характер.

1.1. Иерархический метод классификации всего множества понятий русского языка позволяет естественным образом отразить их семантическую многомерность. Основой

семантической классификации слов любого естественного языка является словарь базисных понятий (около 20000), которые нельзя выразить через другие более простые понятия [6, 101-165].

Фрагмент семантического классификатора:

Класс	Характеристика	Пример	Кол-во
$12	Физический Объект	МАТЕРИЯ ОБЪЕКТ ПРЕДМЕТ	64
$123	ФО Поселения	ПОСЕЛЕНИЕ СЕКТОР	40
$1230	ФО Поселения Территория	АРЕАЛ БАССЕЙН ВОТЧИНА РАЙОН	50
$12311	ФО Поселения Страны Уезд	АНКЛАВ КАНТОН ОБЛАСТЬ ШТАТ	83
$12314	ФО Поселения Страны Город	ГОРОД СТОЛИЦА ПОРТ	79

На основании данного классификатора можно построить иерархическую систему понятий для любого естественного языка.

Узлы семантического дерева понятий двух языков, за редким исключением, не будут совпадать, поскольку сложно найти два слова из разных языков, которые представляли бы два идеально совпадающих понятия. Структура же семантического класса остается постоянной.

1.2. Все слова, являющиеся производными от базисных понятий, описываются функциональной конструкцией, являющейся суперпозицией базисных функций и базисных понятий.

ВНУТРИГОРОДСКОЙ N%~ГОРОД$12314 (Loc(НЕЧТО$1,ВНУТРИ$12/313/05(ГОРОД$12314)))

Описание объектов с помощью конструирующей их функции соответствует функциональной модели описания искусственных языков [7]. Данный подход отражает функциональную природу естественного языка и позволяет построить модель, адекватную этому языку.

Описание английской лексемы может отличаться от описания ее русского перевода. Понятие, которое отражает английская лексема, аппроксимируется целым набором семантических описаний русских лексем. Не обязательно все описания будут соответствовать одному семантическому классу.

1.3. Несмотря на то, что иерархическая система классификации близка к естественному представлению человека об окружающем мире, природа слов такова, что они часто тяготеют к различным ветвям семантического дерева.

Отчасти причиной является то, что когда перед человеком открывается новый мир понятий, он склонен переносить на него «старые» имена, описывающие схожую систему объектов, вместо того, чтобы придумывать «новые» имена [5].

2. Семантические отношения

Существует необходимость в описании семантической близости объектов, принадлежащих разным классам. Синонимические отношения в некоторой степени отражают семантическую близость слов. Одним из возможных решений задачи определения семантической близости является использование таких критериев, как расстояние Левенштейна [2].

При рассмотрении отношений между различными семантическими классами можно учитывать связи, которые представлены в двуязычных словарях.

Например, английская лексема *speculation* может быть отнесена к различным классам, которые соответствуют возможным ее переводам на русский язык.

$13154 обдумывание (calculation, cogitation, deliberation, reflection);

$13154 размышление (reflection, thought, meditation);

$13154 мысли (thoughts, ideas);

$13156 предположение (supposition, assumption, guess-work);

$131241 спекуляция (profiteering, jobbery, gamble);

$14213 теория (theory);

$1440205 слухи (rumor, hearsay, news, sigh, buzz);

$14461 прогноз (forecast, prognosis);

$14461 догадка (guess, conjecture, surmise, guess-work).

Этот пример отражает возможные семантические связи между классами. Особое внимание в данных синонимических отношениях необходимо обратить на синтагматические свойства слов [9]. Поэтому в качестве основного критерия при установлении семантической близости будем рассматривать способность различных лексем употребляться в одних и тех же контекстах.

Семантические отношения между классами позволяют определить русскоязычные понятия, отражающие схожие явления, описываемые английским словом в данном конкретном предложении.

2.1. Языковое отражение окружающего мира зависит от традиций и культуры народа – носителя языка. «Языки сильно различаются и по категориям, которые они выражают, и по конкретным языковым средствам, используемым для выражения этих категорий. Эти различия идут гораздо глубже такого известного факта, что большинство слов не имеет абсолютно точных переводных эквивалентов в других языках, потому что, как вы видели из приведенных примеров, категории могут выражаться как лексически, так и грамматически» [3, 126]. Языковая относительность понятий проявляется в различии выделяемых объектов и их свойств.

Принадлежность объекта классу характеризуется определенным набором свойств этого объекта. Эти свойства в базе знаний принимают вид атрибутов объекта. Набор атрибутов определен для каждого семантического класса.

Лексема, соответствующая конкретному понятию, отнесена к тому классу, на основании свойств которого будет конструироваться объект. Семантическое описание лексемы отражает характер взаимодействия с объектами других классов.

2.2. Отношения, в которые могут вступать объекты, соответствующие лексемам одного предложения, зависят от свойств объектов. Свойства объектов определяются как классом конструирующей его лексемы, так и семантическим описанием.

Отношения между семантическими классами могут описываться как предложно-падежными формами, так и базисными семантическими функциями, а также путем явного указания класса в семантическом описании лексемы. Отношения проявляющиеся на синтаксическом уровне могут носить семантический характер.

Например, для класса $10001 ВЗГЛЯД, куда входят лексемы: *взгляд, бросать, взглянуть, взирать, глядеть, смотреть, поглядывать* тип связи !Куда\!наВин, которая присутствует в описании этих лексем, характеризует отношение, которое имеет то же значение, что и связь !at в описании английских лексем: *look, glance, gaze, stare, glare, cast, goggle, skew, frown*. Для класса $1/215 ГОДНОСТЬ связь !наВин соответствует типу связи !for. В классе $1/27 НАВЛЕКАТЬ связь !наВин соответствует !on (*incur, bring, draw*).

Различные свойства класса могут быть описаны одинаковыми функциональными конструкциями, но будут иметь свою уникальную семантику. Это значение может быть вычислено, в результате инициализации аргументов функции.

Так базисные семантические формулы отражают отношения, значения которых можно вычислить на основании аргументов этих функций. Например, функция Hab [8]:

Hab (НЕЧТО$1, ЦВЕТ$12/012)	ЦВЕТ$12/012	\Красный цветок
Hab (ЧЕЛОВЕК$1241, ВЕЩЬ$1213)	СОБСТВЕННОСТЬ$12411/0	\Дом Ивана
Hab (ДОКУМЕНТ$14002, ИНФОРМАЦИЯ$1440)	СОДЕРЖАНИЕ$1440200	\Книга анекдотов
Hab(ПОСУДА$12131111, ЖИДКОСТЬ$121112)	СОДЕРЖИМОЕ$12/013	\Бутылка молока

2.3. Часть атрибутов объекта входит в семантическое описание лексемы. Они образуют, так называемые, точные семантические связи. Свободные семантические связи и чисто синтаксические связи не входят в семантическое описание, но учитываются анализатором.

3. Вычисление значения английской лексемы

Результат вычисления значения английской лексемы зависит от контекста. То, какой объект будет сконструирован в результате семантического анализа, определяется

объектами, с которыми он будет взаимодействовать. Семантическая функция, описывающая английскую лексему, позволяет вычислить объект и его свойства, который будет достаточно точно указывать на русскую лексему.

3.1. Для описания английской лексемы могут использоваться семантические функции соответствующих русских лексем. Но аргументы этих функций должны содержать типы связей, отражающие характер отношений, принятый в английском языке. Поэтому аргументы английских лексем будут отличаться предложно-падежной формой аргумента, а в ряде случаев и классом [1].

FALL N%~СУДЬБА$112(**ДОЛЯ$12/013141~!to**, КОМУ:**!to**, **!Инфин**)
= ВЫПАДАТЬ \Some painful moments may fall to her lot; the hardest work fell to her share; it fell to me to break the news to her.

FALL N%~ВЫПАДЕНИЕ$15428(**СНЕГ$122156\ЗУБ$104113\ВОЛОС$10412~!Им**, **НЕЧТО$1~!Откуда**)
= ВЫПАДАТЬ \A child's first teeth fall; his hair is falling; snow fell.

FALL N%~МЕСТОНАХОЖДЕНИЕ$12/2(**!Им**, **ЗАПАДНЯ$121312\РУКА$1043~!into**)
= ПОПАДАТЬ

FALL N%~ВПАДЕНИЕ$154201(**РЕКА$122426~!Им**, **МОРЕ$12101~!Куда**)
= ВПАДАТЬ \Rivers that fall into the sea

FALL N%~ПСИХИКА$13(**ЧЕЛОВЕК$1241~!Им**, **ПСИХИКА$13~!Куда**)
= ВПАДАТЬ \To fall into disgrace; to fall into a reverie; to fall into rage.

FALL N%~ТИХИЙ$12/00601(**ЗВУК$12/006\ВЕТЕР$12211\ПЛАМЯ$12/01312~!Им**)
= ЗАТИХАТЬ \The wind fell; the flames rose and fell; here his voice fell.

FALL N%~ОСВЕЩЕНИЕ$12/007(**СВЕТ$12/007~!Им**, **НЕЧТО$1~!Куда\!on**)
= ПАДАТЬ

FALL N%~ЗАВИСИМОСТЬ$131263(!Им, **ЗАВИСИМОСТЬ$1126~!under**)
= ПОПАДАТЬ

FALL N%~ВРЕМЯ$160(**СОБЫТИЕ$11\ДЕЙСТВИЕ$15~!Им**, **СОБЫТИЕ$11\ВРЕМЯ$16~!on**)
= ПРИХОДИТЬСЯ

В данном подходе для вычисления перевода английской лексемы используются только аргументы семантической функции. Они могут достаточно точно определять русскую лексему. Описание английской лексемы представлено несколькими альтернативами, что приводит к некоторому увеличению словаря. На самом деле в словаре будут присутствовать английские лексемы с одинаковым описанием. Поэтому одно описание можно сопоставлять нескольким лексемам.

GET, FALL N%~МЕСТОНАХОЖДЕНИЕ$12/2(**!Им**, **ЗАПАДНЯ$121312\НЕПРИЯТНОСТЬ$1303\РУКА$1043~!into**)
= ПОПАДАТЬ \ Otherwise, you can get into a trap. You can fall into a trap and with their offensive line you're going to pay for it.

3.2. В общем случае английской лексеме, как и русской, соответствует семантическая функция, которая не обязательно совпадает с описанием ее перевода на русский язык.

ОСТРИЕ $12/105(S1>Copul(S1:ЧАСТЬ$12/013141(!Род),ОСТРЫЙ$12/105))

SPIKE $12/105(S1>Copul(S1:ЧАСТЬ$12/013141(!of\!Род),Rel(ОСТРЫЙ$12/105,МЕТАЛЛ$12122))

Данная функция может принимать следующие значения:

1) предмет, имеющий металлическое острие: шип, гвоздь, костыль (для скрепления рельсов), клин, зубец, каблук "шпилька", штырь, наколка (для накалывания на чеков), шприц для подкожных инъекций, горные ботинки, шиповки, беговые туфли;

2) предмет подобный острию: молодая скумбрия, неразветвленный рог молодого оленя, колос;

3) процесс: перепад напряжения (в электрической сети), резкий скачок цен.

Можно дифференцировать значения лексемы, сопоставив ей различные семантические описания.

SPIKE N%~ОСТРЫЙ$12/105(Caus(!Им,IncepMult_Hab(!Вин,ОСТРЫЙ$12/105(!with\!Тв))) [[**СНАБЖАТЬ**$20]])

SPIKE N%~ОСТРЫЙ$12/105(PerfCaus(!Им,IncepMult_Hab(РЕЧЬ$14402~!Вин, ОСТРЫЙ$12/105(!with\!Тв))) [[**ОЖИВИТЬ**$20]])

SPIKE N%~ОСТРЫЙ$12/105(Caus(!Им,IncepLab(МОЛВА$1440205~!Вин, Mult(ОСТРЫЙ$12/105))) [[**ОПРОВЕРГАТЬ**$20]])

SPIKE N%~ОСТРЫЙ$12/105(PerfCaus(!Им,IncepLab(ПЛАН$1442\ЗАМЫСЕЛ$13154~!Вин, Mult(ОСТРЫЙ$12/105))) [[**РАССТРОИТЬ**$20]]) \\пригвоздить

SPIKE N%~ОСТРЫЙ$12/105(PerfCaus(ПРОЦЕНТ$12/0311~!by\!наВин, IncepCopul(ПРИЕМ$15429\ПРОДАЖА$152822\ОБЪЕМ$12/016\ИНФЛЯЦИЯ$12146~!Им,ОСТРЫЙ$12/105)) [[**РЕЗКО_УВЕЛИЧИТЬСЯ**$20]])

1) снабжать остриями (делать что-либо обладателем остриев): прибивать гвоздями, прокалывать (не вытаскивая обратно), поранить шипами (оставить след), нанизывать накалывать.

2) коннотационное значение: приводить в негодность (оборудование); пресекать, опровергать (слухи, молву); отвергнуть (статью, наколов ее на штырь); добавить алкоголя в напиток; оживить (речь); сдобрить (суп перцем); ввести токсичное вещество в организм (с помощью инъекции).

3.3. В этом случае семантическое описание лексемы будет относиться к определенному классу. По данному семантическому описанию на основании имеющихся

отношений между классами можно построить объект другого класса, который наиболее точно описывается русской лексемой этого класса.

3.4. Указывая область определения семантической функции, мы устанавливаем отношения между классами. Эти отношения можно использовать для вычисления перевода английских лексем.

RENEW (ДЕЙСТВИЕ$15) [[ВОЗОБНОВИТЬ$20]]
RENEW (ОДЕЖДА$12136) [[ОБНОВЛЯТЬ$20]]
RENEW (КЛЯТВА$1440212) [[ПОВТОРЯТЬ$20]]
RENEW (ЧУВСТВО$1300) [[ВЫЗЫВАТЬ$20]]
RENEW (СИЛЫ$12411202\ОТНОШЕНИЯ$1/3232) [[ВОССТАНАВЛИВАТЬ$20]]
RENEW (ЗАПАСЫ$12/016) [[ПОПОЛНЯТЬ$20]]
RENEW (ПОЛИС$144902\СРОК$16132) [[ПРОДЛЕВАТЬ$20]]
RENEW (РЕЧЬ$14402) [[ПОРОДОЖАТЬ$20]]

3.5. Отношения между классами можно выявлять, исходя из структуры конкретного объекта. Аргументы объекта будут указывать на класс, объект которого должен быть построен. Это важно, когда связь носит чисто синтаксический характер. Например, при вычислении значений прилагательных.

БОГАТЫЙ $12411/202 (A1>Hab(A1:НЕЧТО$1,ДОСТАТОК$1/21062(!Тв\!вПред\!наВин,ПРИЧИНА:!Ото)))
RICH $12411/202 (A1>Hab(A1:НЕЧТО$1,ДОСТАТОК$1/21062(!in,ПРИЧИНА:!from)))

В описании прилагательных не присутствуют аргументы, описывающие отношения с классом определяемого слова. В то время как существуют отношения:

RICH (СЛИВКИ$10138\ЛИСТВА$1223/2) [[ГУСТОЙ$12110/02]]
RICH (МОЛОКО$10138) [[ЖИРНЫЙ$121111]]
RICH (БЛЮДО$1013) [[ПИТАТЕЛЬНЫЙ$1001]]
RICH (ПОЧВА$121252) [[ПЛОДОРОДНЫЙ$12233]]
RICH (ФРУКТ$122332) [[СОЧНЫЙ10112]]
RICH (ПЕЙЗАЖ$141312) [[КРАСИВЫЙ$12/020315]]
RICH (ТЕМА$1440200) [[НЕИСЧЕРПАЕМЫЙ$15427]]
RICH (ЧЕЛОВЕК$1241) [[БОГАТЫЙ$12411/202]]
RICH (ТОН$12/012) [[ГЛУБОКИЙ$12/01403]]
RICH (ПЛАТЬЕ$121363) [[ДОРОГОЙ$1/214]]
RICH (НАГРАДА$12411/031\ ПРЕДЛОЖЕНИЕ $1440219) [[ЦЕННЫЙ$1/214]]

Объекты, определяющие значение функции RICH(X), имеют свойство, которое отражает ценную характеристику. Для множественных объектов – это их количество (ГУСТОТА, КОНЦЕНТРАЦИЯ), для пищи – питательные элементы (СОК, ЖИР), для вещей, имеющих стоимостное выражение – (ЦЕНА). В большинстве остальных случаев в качестве перевода можно выбрать основное значение слова RICH – БОГАТЫЙ.

3.6. Если невозможно определить русское значение английского слова, основываясь только на классе объекта, необходимо использовать атрибуты этого объекта. Для этого используется следующее семантическое описание лексемы RICH:

RICH (A1>Hab(A1:НЕЧТО$1,Magn(**СВОЙСТВО**:!in)))[[БОГАТЫЙ$20]]

Значение аргумента СВОЙСТВО ссылается на соответствующий класс, объект которого должен быть построен. Вычисление лексемы будет опираться на эквивалентное семантическое описание. Использование эквивалентных преобразований семантических формул позволяет определить следующие значения слова RICH:

ЖИРНЫЙ (A1> Content(A1:НЕЧТО$1,ЖИР$121111))

ГУСТОЙ (A1>Hab(A1:НЕЧТО$1,Magn(КОНЦЕНТРАЦИЯ$14216/1)))

СОЧНЫЙ (A1>Hab(A1:НЕЧТО$1,СОК$10112))

БОГАТЫЙ (A1>Hab(A1:НЕЧТО$1,ДОСТАТОК$1/21062))

ГЛУБОКИЙ (A1>Hab(A1:НЕЧТО$1,Magn(ГЛУБИНА$12/01403)))

ДОРОГОЙ (A1>Hab(A1:НЕЧТО$1,Magn(ЦЕНА$1/214)))

ЦЕННЫЙ (A1>Hab(A1:НЕЧТО$1,Magn(ЦЕНА$1/214)))

Некоторые эквивалентные семантические преобразование могут выполняться для ограниченного ряда классов.

ПЛОДОРОДНЫЙ (A1>EmCaus(A1:НЕЧТО$1,IncepFunc(Mult(ПЛОД$12233))))

ПИТАТЕЛЬНЫЙ (A1>EmCaus(Content(A1:НЕЧТО$1,Z1),Hab(Z2,ПИЩА$101)))

Этот подход оказывается эффективным, если взаимодействия происходят не на уровне классов, а на уровне объектов (МОЛОКО$10138\ СЛИВКИ$10138).

Заключение

При вычислении значения английских лексем целесообразно использовать различные способы определения отношений между семантическими классами и их объектами. Семантические связи между объектами реализуются как на уровне семантического описания, так и на уровне классификатора и самого алгоритма семантического анализа.

Для корректного перевода необходимо использовать и возможности базы знаний, которая должна отражать свойства объектов в разрезе, как русского языка, так и английского, тем самым, позволяя вычислить семантику слова любой сложности.

Результаты представленные в статье доказали свою ценность при составление двуязычного семантического словаря.

Литература

1. Богдановский А.Е. Семантический анализ текстов на английском языке // Процессы управления и устойчивость: Тр. 37-й междунар. науч. конф. аспирантов и студентов. СПб., 10–13 апреля 2006 г. / Под ред. А. В. Платонова, Н. В. Смирнова. – СПб.: Изд-во С.-Петерб. ун-та, 2006. – с. 286–293.

2. Мозговой М.В. Контекстно-ориентированный тезаурус русского языка // Процессы управления и устойчивость: Тр. 37-й междунар. науч. конф. аспирантов и студентов. СПб., 10–13 апреля 2006 г. / Под ред. А. В. Платонова, Н. В. Смирнова. – СПб.: Изд-во С.-Петерб. ун-та, 2006. – с. 379–383.
3. Слобин Д., Грин Дж. Психолингвистика. Перевод с английского Е. И. Негневицкой/ Под общей редакцией и с предисловием доктора филологических наук А. А. Леонтьева. – М.: Прогресс, 1976. – 336 с. С.125-138
4. Социология: Энциклопедия / Сост. А.А. Грицанов, В.Л. Абушенко, Г.М. Евелькин, Г.Н. Соколова, О.В. Терещенко. – Мн.: Книжный Дом, 2003. – 1312 с.
5. Степанов Ю. С. Семиотика: Антология. – М.: Академический проект, – 2001.– 702 с.
6. Тузов В.А. Компьютерная семантика русского языка. – СПб.: Изд-во СПбГУ, 2004. – 400 с.
7. Тузов В.А. Математическая модель языка. – Л.: Изд-во Ленингр. Ун-та, 1984. – 176 с.
8. Тузов В.А. Семантика предложно-падежных форм русского языка // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.) / Под ред. Н.И. Лауфер, А. С. Нариньяни, В. П. Селегея. – М.: Изд-во РГГУ, 2006. С. 513-518.
9. Черняк В.Д. Проблема синонимии и лексико-грамматическая классификация слов [Электронный ресурс] : [в т. ч. анализ деривацион. процессов] // Новосибирский государственный педагогический университет : интернет-портал. – Новосибирск, 2001 <http://www.nspu.net/fileadmin/library/books/2/web/xrest/article/leksika/sinonim/che_art01.htm>.