АВТОМАТИЧЕСКИЙ ПЕРЕВОД СЕМАНТИЧЕСКОЙ СЕТИ WORDNET НА РУССКИЙ ЯЗЫК

И.Г. Гельфенбейн
Новосибирский Государственный Университет
Россия, 630090, Новосибирск, Пирогова, 2
e-mail: ilya@gorodok.net

А.В. Гончарук
Институт математики СО РАН
Россия, 630090, Новосибирск, пр. ак. Коптюга, 4
e-mail: artemgoncharuk@ngs.ru

В.П. Лехельт
Новосибирский Государственный Университет
Россия, 630090, Новосибирск, Пирогова, 2
e-mail: vlt@ngs.ru

А.А. Липатов
Новосибирский Государственный Университет
Россия, 630090, Новосибирск, Пирогова, 2
e-mail: alipatov@gorodok.net

В.В. Шило
Новосибирский Государственный Университет
Россия, 630090, Новосибирск, Пирогова, 2
e-mail: vitay@inet.ssc.nsu.ru

1. Введение

В настоящее время достаточно популярными становятся электронные тезаурусы, схожие по своей структуре с лексическо-семантической базой данных WordNet (http://www.cogsci.princeton.edu/~wn/), разработанной в Принстонском университете (США) в 1985 году группой ученых под руководством Дж. Миллера. WordNet состоит из четырех частей, содержащих отдельно существительные, глаголы, прилагательные и наречия. Каждая часть представляет собой семантическую сеть, узлами которой являются синонимические ряды (синсеты) соответствующей части речи, отражающие смыслы понятий.

Существует достаточно большое количество WordNet-подобных баз данных для различных языков, например, под эгидой проекта EuroWordNet (http://www.illc.uva.nl/EuroWordNet/) были объединены лексико-семантические базы данных для европейских языков. Главной особенностью EuroWordNet является выделение общей понятийной части, так называемой Top Ontology, и межъязыковых индексов (Inter-Lingual-Index), которые сопоставляют между собой понятия разных языков. Также существует Всемирная Организация WordNet (Global WordNet Association – http://www.globalwordnet.org/), объединяющая большое число разработчиков подобных систем по всему миру.

Разработка подобных систем для русского языка ведется несколькими исследовательскими группами:

- RussNet в СпбГУ, кафедра математической лингвистики (http://www.phil.pu.ru/)
- Центр информационных технологий МГУ

Построение базы данных в проекте RussNet производится вручную, что позволяет получить качественный тезаурус, учитывающий специфические особенности русского языка.

В настоящей работе рассматривается подход, позволяющий частично автоматизировать процесс получения WordNet-подобной базы данных для русского языка. Основная идея состоит в замене английских синсетов на их русские аналоги при сохранении структуры семантической сети. В некоторых случаях такую замену можно произвести автоматически, учитывая то, что замене подлежат не произвольные множества слов и словосочетаний, а множества слов, связанных отношением синонимии. В качестве базы для процедуры перевода используется English Princeton WordNet. Работа содержит описание результатов применения такого подхода и анализ случаев, в которых целесообразно и нецелесообразно применение такого подхода.

Результат данной работы может использоваться в различных интеллектуальных системах, переводчиках, поисковых системах и т.п. В перспективе планируется интеграция баз русского WordNet'а в структуру EuroWordNet.

Работа выполнена в рамках стажировки студентов Новосибирского Государственного Университета в компании Novosoft Inc (http://www.novosoft-usa.com/, http://www.novosoft.ru/).

2. Алгоритм преобразования семантической сети.

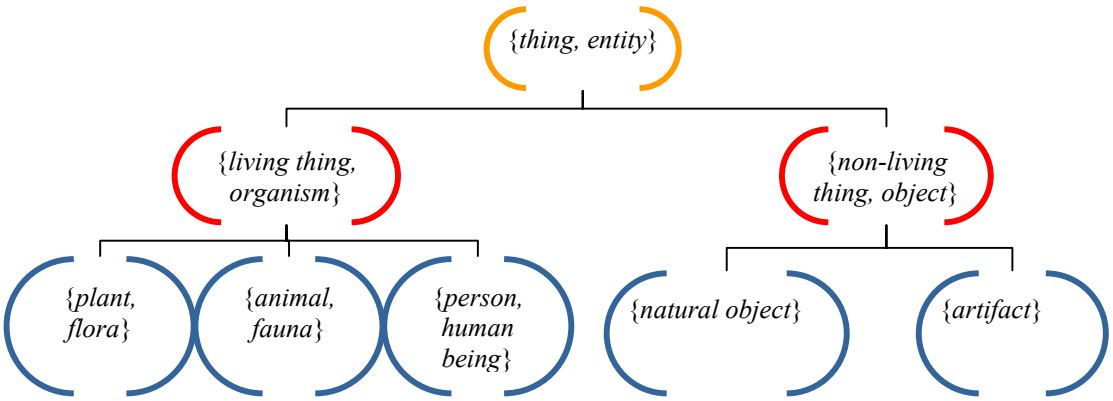
2.1 Постановка задачи

Для начала небольшое введение в предметную область.

Синсет – синонимический ряд – множество слов, связанных отношением синонимии, являющимся разбиением множества всех лексических единиц на классы эквивалентности, выражающие сущность каких-либо понятий.

Примеры синсетов: *{good, fine}, {man, adult male}*.

WordNet – семантическая сеть, в узлах которой находятся синсеты, связанные различными отношениями, такими как гипонимия, гиперонимия, голонимия, меронимия и т.д. Каждый синсет имеет описание на естественном языке и примеры использования входящих в него слов.



Графическое представление семантической сети WordNet

Исходя из предположения, что сама семантическая сеть смыслов не зависит от языка ее представления, заменяем английские синсеты на русские, сохраняя все отношения между ними. В итоге получается сеть, сохраняющая исходную структуру, изменяются лишь описания смыслов – теперь они представляются на другом естественном языке (русском). Однако выполнить полное автоматическое преобразование сети невозможно вследствие того, что для некоторых английских синсетов, описывающих особые смыслы, характерные для английского языка, не существует эквивалентов в русском языке. Такие синсеты можно будет просто удалить.

Итак, задача алгоритма – попытка найти русские синсеты, выражающие наиболее близко смысл английских. Описания и примеры употребления для синсетов пока не переводятся.

Будем обозначать отдельные слова маленькими латинскими буквами, множества слов большими латинскими буквами, наборы таких множеств – словами.
Пример: $w \in S \in Synsets$.

В нашем распоряжении имеются:

- **Англо-русский словарь (English-to-Russian Dictionary – *ERD*).** Каждому английскому слову из словаря сопоставляется словарная статья, содержащая русские слова (переводы), сгруппированные по смыслу. Эти группы слов рассортированы по употребительности, начиная с самого часто используемого.

“be” $\xrightarrow{ERD}$ *({быть, существовать}, {находиться})*

- **Синонимический словарь (Synonymy Dictionary – *SD*).** Представляет собой множество синсетов на русском языке.
- **Частотный словарь (English Frequency List – *EFL*).** Каждому английскому слову из словаря сопоставляется некоторое натуральное число (1, 2, ...), характеризующее его частотность, т.е. количество вхождений в некоторый фиксированный корпус текстов. *EFD(“the”) = 6187267, EFD(“time”) = 183427.*

Для существительных и глаголов преобразование сети лучше проводить, начиная с корней дерева гипонимии (для существительных) или тропонимии

(для глаголов) и далее по дереву. Это нужно для отсечения непереводимых синсетов и их потомков. Пока это не реализовано.

2.2 Алгоритм

Пусть имеется синсет $\{w_1, w_2, ..., w_n\}$, состоящий из n слов w_i. В зависимости от значения n возможно несколько вариантов алгоритма:

1. $n \geq 2$

Переводим с помощью англо-русского словаря каждое слово из английского синсета и получаем множество словарных статей: $w_i \xrightarrow{ERD} \{R_i^1, ..., R_i^{k_i}\}$. Слова каждой статьи объединяем в «множество переводов английского слова»: $R_i = \bigcup_{j=1}^{k_i} R_i^j$.

Обозначим: $I = \bigcap_{i=1}^{n} R_i$ - множество слов, входящих во все рассматриваемые словарные статьи.

Из синонимического словаря выбираем русские синсеты, содержащие хотя бы один перевод слова из английского синсета.

$$Synsets = \{S \in SD \mid \exists i : S \cap R_i \neq \varnothing\}$$

Если это множество пусто, значит, слова из данного английского синсета просто не содержатся в англо-русском словаре. В таком случае перевод синсета невозможен. Поэтому далее полагаем, что множество *Synsets* не пусто.

Определим вес синсета в зависимости от количества слов в множестве I и рассмотрим два случая:

Случай 1.1. $|I| \geq 1$, т.е. существует хотя бы одно слово, входящее во все словарные статьи. Это означает, что все переводимые английские слова, вероятно, имеют какой-то общий смысл.

Для каждого синсета S из множества *Synsets* находим количество слов, которые содержатся и в пересечении переводов I и в синсете:

$$weight(S) = |S \cap I|.$$

Назовем это **весом** синсета.

Случай 1.2. $|I| = 0$, т.е. пересечение пусто.

Для каждого слова r из множества всех переводов английских слов находим количество его повторных вхождений в синсеты из *Synsets*:

$$NumberOfRepetitions(r) = |\{S \in Synsets \mid r \in S\}| - 1.$$

Назовем это числом повторов.

В этом случае вес синсета будем считать как сумму числа повторов для каждого слова, входящего в синсет:

$$weight(S) = \sum_{r \in S} NumberOfRepetitions(r).$$

Далее, после того, как определен вес синсета, находим из множества *Synsets* все синсеты с максимальным весом (*maxweight*):

$$SelectedSynsets = \{S \in Synsets \mid weight(S) = maxweight\}.$$

Данное множество не может быть пустым, т.к. множество *Synsets* не пусто. Если в *SelectedSynsets* содержится единственный синсет – он и будет искомым. Если же там синсетов несколько, то необходимо выбрать наиболее подходящий. Для этого могут быть использованы различные алгоритмы, одним из них является выбор при помощи частотного словаря. Т.е. полагаем, что если русский синсет содержит переводы достаточно редких слов, то он и является искомым, т.к. редкие слова и указывают наиболее точный смысл переводимого английского синсета.

Сначала для всех слов из синсетов множества *SelectedSynsets* определим **минимальную частоту исходного слова**:

$$Frequency(r) = \min_{w_i : r \in R_i} EFD(w_i).$$

Теперь для всех синсетов из *SelectedSynsets* определим **среднюю частоту исходных слов**.

$$AverageFrequency(S) = \frac{\sum_{r \in S} Frequency(r)}{|S|}.$$

Далее считаем, что искомым является синсет из множества *SelectedSynsets*, который имеет *наименьшую* среднюю частоту исходных слов.

2. $n = 1$, то есть синсет состоит всего из одного слова.

Это наиболее простой случай, т.к. здесь можно использовать уже готовое упорядочивание групп переводов по употребительности в англо-русском словаре. Т.е. выбирается синсет, наиболее близкий первому множеству слов, содержащемуся в переводах слова из английского однословного синсета. Под близостью подразумевается, во-первых, максимальное пересечение и, во-вторых, минимальное количество слов, на которых они отличаются.

Примеры перевода синсетов:

Случай 1 ($n \geq 2$):

1.1 ($|I| \geq 1$):
Исходный английский синсет: **{detent, dog, click, pawl}**
a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward

Множества переводов:
R(detent) = {защёлка, собачка, ...}
R(dog) = {собачка, собака, "свинья", пёс, самец_волка, кобель, ...}
R(click) = {защёлка, собачка, щёлканье, щелчок, ...}
R(pawl) = {пал, предохранитель, собачка}

I **=** {собачка}, т.е. все синсеты пересеклись по одному слову
SelectedSynsets = {{собачка, защёлка}, {собачка}, {стопор, собачка, защёлка}}

Используя алгоритм работы с частотным словарем, выбираем вариант **{стопор, собачка, защёлка}**

1.2 ($|I| = 0$):
Исходный английский синсет: **{motorcar, car, machine, automobile, auto}**
4-wheeled motor vehicle; usually propelled by an internal combustion engine; "he needs a car to get to work"

Множества переводов:
R(motorcar) = {}
R(car) = {повозка, машина, кабина_лифта, колесница, автомобиль, тележка,...}
R(machine) = {швейная_машинка, механизм, машина, автомобиль, аппарат, самолёт, станок}
R(automobile) = {автомобиль}
R(auto) = {автомобиль}

I**=**{}, в данном случае не нашлось слов, содержащихся во всех синсетах
NumberOfRepetitions:{вагонетка=0, машина=1, автомобиль=2, колесница=0, ...} – отображает слово и количество его повторений во всех синсетах
SelectedSynsets = {{автомобиль, машина}}

Множество SelectedSynsets содержит только один синсет, и поэтому, без использования частотного словаря, выбираем вариант **{автомобиль, машина}**

Случай 2 ($n = 1$):
Исходный английский синсет: **{good}** - *существительное*
benefit; "for your own good"; "what's the good of worrying?"

Первый перевод в словаре: **{добро, благо}**

2.3 Используемые ресурсы и реализация

В качестве исходной семантической сети используем Princeton WordNet 1.7.1.

Используем следующие словари:

1. ***Англо-русский словарь*** *– как можно более полный словарь, в котором слова, описывающие различные понятия, являющиеся переводами слова, разбиты по группам, а группы упорядочены по употребительности.*

В качестве англо-русского словаря используем словарь Мюллера. Это связано с тем, что он содержит достаточно большой объем лексики и имеется в виде неплохо формализованного текстового файла в формате MOVA, распространяемого по лицензии GPL (General Public License). Словарь преобразуется из формата MOVA в XML-формат для удобства использования и лучшей структуризации. Недостатком словаря Мюллера является отсутствие современной лексики. Кроме того есть недостатки у его текстовой версии в MOVA-формате – ошибки в формализованной структуре словаря.

Пример словарной статьи:

car [kA:] _n. 1> автомобиль, машина 2> вагон (трамвая, _ам. тж. железнодорожный); parlor car _ам. салонвагон; hand car дрезина 3> тележка; повозка, вагонетка 4> гондола дирижабля 5> _ам. кабина лифта 6> _поэт. Колесница

2. ***Синонимический словарь*** *– словарь, представляющий собой множество синонимических рядов, каждый из которых описывает некоторое понятие. Под синонимией здесь подразумеваем синонимию WordNet'а. Чем больше множество описываемых понятий, тем словарь лучше.*

Сначала пытались использовать синонимический словарь под ред. Евгеньевой, словарь синонимов из системы ЭТАП. Но эти словари содержат синонимию, которая не подходит для целей преобразования. Поэтому используемый словарь создается из англо-русского словаря Мюллера выделением из него синонимических рядов, описывающих отдельные понятия.

Пример:

{автомобиль, машина}, {вагон}, {тележка, повозка, вагонетка}, {кабина лифта}, {колесница }

3. ***Частотный словарь*** *– словарь, сопоставляющий различным английским словам число их вхождений в некоторый корпус текстов.*

Используется частотный словарь Адама Килгаррифа в формате British National Corpus (http://www.itri.bton.ac.uk/~Adam.Kilgarriff/bnc-readme.html)

Проект реализован на языке программирования Java (JDK 1.3.1) (http://java.sun.com/), языка разметки XML (http://www.w3.org/XML/), библиотек SAX (http://www.saxproject.org/) и Xerces 2 (http://xml.apache.org/xerces2-j/).

Для представления полученных баз при проверке используется утилита WordNet TreeWalk 1.8.1 (http://www.ac-toulouse.fr/wordnet/).

2.4 Возможности проверки и пополнения полученной базы

После генерации баз русского WordNet'а необходима проверка корректности сопоставления русских и английских синсетов. Для поиска ошибок в полученной семантической сети используется выборка объема ~200 синсетов (0,2% от общего числа). Выборка создается так, чтобы наиболее полно охватить различные виды содержащихся в WordNet'е синсетов. При ее составлении учитывается часть речи, размер синсета, наличие и виды связей, положение синсета в дереве гипонимии и т.п.

Проверка производится следующим образом: сравниваются смыслы исходного английского синсета и полученного русского. Для их уточнения используется описание исходного английского синсета на естественном языке, его связи с другими синсетами в семантической сети. Также необходимо проверять правильность связей между получившимися русскими синсетами.

Для проверки, доработки и пополнения полученных баз данных русского WordNet'а разрабатывается приложение (RussianWordNet EditTool), при помощи которого можно будет вносить изменения в уже имеющуюся структуру баз: добавлять/удалять синсеты, их описания, связи между ними и т.п.

3. Заключение.

С помощью вышеописанного алгоритма было переведено около 45% английской семантической сети. Из них приблизительно 75% корректно. В числе непереведенных синсетов - сленг, выражения, смысл которых трудно описать в русском языке, и просто синсеты, в которых ни одно слово не содержится в использовавшемся англо-русском словаре.

В будущем планируется более тщательная проверка полученных результатов, удаление неправильных переводов, пополнение сети новыми синсетами и лингвистическими отношениями.

Описанная методика применима для генерации сетей типа WordNet не только для русского языка, но и для многих других.

4. Список литературы.

1) Miller, G. et al (1993). Five Papers on WordNet. Technical Report, Cognitive Science Laboratory, Princeton University.
2) Словарь синонимов русского языка./ под ред А. П. Евгеньева. Л., 1970.
3) Апресян Ю.Д. Новый объяснительный словарь синонимов русского языка. М., 1997.
4) Мюллер В.К. и др. Англо-русский словарь. М., 1999