# АВТОМАТИЗАЦИЯ ПРОЦЕССА ПОСТРОЕНИЯ И ПОПОЛНЕНИЯ ДВУЯЗЫЧНЫХ СПЕЦИАЛИЗИРОВАННЫХ СЛОВАРЕЙ[1]

# AUTOMATION OF THE PROCESS OF CREATING AND WIDENING OF BILINGUAL SPECIALIZED DICTIONARIES

***А.А. Липатов***

***Новосибирский государственный университет***

***anton.lipatov@gmail.com***

***А.А. Мальцев***

***Институт математики СО РАН***

***amalcev@mail.ru***

***В.В. Шило***

***Новосибирский государственный университет***

***victor.shilo@gmail.com***

В настоящей работе рассматривается подход, позволяющий автоматизировать процесс построения специализированного двуязычного словаря для фиксированной предметной области. Для решения задачи используется уже затраченный труд переводчиков. Для построения используются тексты по нужным тематикам на английском языке и их переводы на русский язык, сделанные человеком. Суть алгоритмов для каждого текста и его перевода состоит в нахождении соответствий между семантическими единицами в парных текстах.

[1] Работа выполнена при финансовой поддержке по проекту 8328 программы Рособразования "Развитие научного потенциала высшей школы"

### *Введение*

В настоящей работе рассматривается подход, позволяющий автоматизировать процесс построения специализированного двуязычного словаря для фиксированной предметной области. В частности, данный подход оказывается полезным при расширении семантических сетей типа WordNet.

Для решения задачи используется уже затраченный труд переводчиков. Для построения используются тексты по нужным тематикам на английском языке и их переводы на русский язык, сделанные человеком. Суть алгоритмов для каждого текста и его перевода состоит в нахождении соответствий между семантическими единицами в парных текстах. Это требует последовательного разбиения и анализа парных текстов.

Большинство созданных алгоритмов для решения задачи не используют особенностей английского и русского языков и, поэтому, принципиально применимы и к другим парам языков.

### *Постановка задачи*

Пусть есть некоторая заданная предметная область. Существует проблема построения (пополнения) специализированного англо-русского словаря для данной предметной области. В качестве источника информации используется корпус текстов по данной тематике на английском языке и их переводы на русский язык.

В данном случае слово словарь, который нужно построить, рассматривается в двух смыслах: во-первых, как обычный англо-русский словарь, во-вторых, как расширения общелексической части семантических сетей WordNet и RussianWordNet, связанные между собой отношением эквивалентности смыслов синсетов на разных языках.

Целью данной работы было изучить возможности решения данной задачи автоматически или полуавтоматически.

### *Используемые ресурсы*

Созданные алгоритмы используют следующие ресурсы:

- Англо-русский словарь общей лексики.
- Семантические сети WordNet и RussianWordNet, связанные между собой отношением эквивалентности смыслов синсетов на разных языках.
- Морфологические модули для английского и русского языков.

Теперь подробнее об этих ресурсах и требованиям к ним.

### *Англо-русский словарь общей лексики.*

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.

### *Электронная лексическая база WordNet*

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.

### *Электронная лексическая база RussianWordNet*

При построении базы данных в проекте RussianWordNet используется 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. Кроме этого, в качестве источников информации используются: англо-русский словарь Мюллера, генерируемый на его основе синонимический словарь, частотный словарь Адама Килгаррифа.

Кроме того, необходимо особо отметить, что при построении базы RussianWordNet было построено отношение эквивалентности смыслов между синсетами WordNet и RussianWordNet. То есть отношение, показывающее, что смысловые значения русского и английского синсетов (почти) совпадают. «Почти» потому, что при построении лексической базы данных RussianWordNet делается предположение, что сама семантическая сеть смыслов не зависит от языка ее представления, и поэтому производится замена английских синсетов на русские с сохранением всех отношений между ними.

### *Общий алгоритм решения*

При решении задачи используется предположение о монотонности перевода, т.е. том, что при переводе текста порядок предложений и абзацев не изменяется. Конечно, это предположение выполнено не всегда. Алгоритмы были разработаны

таким образом, что в этих исключительных случаях куски текста, где изменился порядок предложений или абзацев при переводе, пропускаются и исключаются из дальнейшего рассмотрения.

Алгоритм решения поставленной задачи состоит из нескольких ступеней. Для каждой из них были разработаны алгоритмы, которые исходят из различных предположений и используют различные ресурсы. Комбинируя данные алгоритмы, можно получить результат, который отвечает поставленным целям наилучшим образом.

Можно выделить три основные части алгоритма, решающие три подзадачи. Каждая часть использует на входе результат работы предыдущей. Первая часть алгоритма применяется последовательно к каждому тексту и его переводу.

### *Разбивка текста на предложения.*

Вход: текст на английском языке и его перевод на русский язык.

Выход: упорядоченные списки предложений из текста на английском языке и из его перевода на русский язык.

Алгоритм последовательно идентифицирует концы предложений. Предложения могут оканчиваться только специфическими знаками препинания (точкой “.”, восклицательным “!” или вопросительным “?” знаками, закрывающимися кавычками “»”, либо переводом строки). При этом учитывается то, что после знака препинания обычно следует один пробел, а затем новое предложение, которое начинается словом с заглавной буквы. Принимаются во внимание также и другие особенности предложений.

### *Сопоставление предложений и переводов (выравнивание текста)*

Вход: упорядоченные списки предложений из текста на английском языке и из его перевода на русский язык.

Выход: множество пар предложений из текста и его перевода.

Для решения данной подзадачи было разработано множество алгоритмов, которые можно применять независимо друг от друга или совместно в некоторой последовательности.

Основная идея этих алгоритмов состоит в следующем. Пусть у нас имеются два упорядоченных списка предложений: *A(1), ..., A(n) и B(1), ..., B(m)*. Допустим, что, используя некоторые соображения, можно сразу сопоставить предложения $A(n_1) \rightarrow B(m_1)$, ..., $A(n_k) \rightarrow B(m_k)$. Причем, исходя из нашего первоначального предположения о монотонности перевода, ясно, что $n_1<...<n_k$ и $m_1<...<m_k$. Тогда на следующем шаге можно перейти к последовательному рассмотрению списков $A(1), ..., A(n_1-1)$ и $B(1), ..., B(m_1-1)$; …; $A(n_k+1), ..., A(n)$ и $B(m_k+1), ..., B(m)$. И так далее продолжаем рекурсивно рассматривать подсписки до тех пор, пока применение алгоритмов не перестанет приносить результатов.

Вот основные алгоритмы для решения данной подзадачи:

1) "Непереведенные слова" Алгоритм осуществляет сопоставление предложений и переводов с одинаковыми словами или группами символов. В результате используются непереведенные слова и выражения (например, латынь или названия компаний на английском языке), числа, имена файлов и т.д.

2) "Однозначно переводимые слова" Существует достаточно большая группа слов, которые переводятся почти однозначно. Например, в этой группе содержатся имена людей, названия и другие имена собственные; аббревиатуры; технические термины; слова из уже имеющегося специализированного словаря по данной предметной области и т.д. Таким образом, если в предложение входит некоторое слово из *словаря однозначно переводимых слов*, в возможный перевод предложения – перевод этого слова, то такие предложения могут быть сопоставлены.

3) "Атрибуты" Алгоритм использует особенности использования специфических знаков препинания для сопоставления предложений. Например, использование восклицательного и вопросительного знаков, знаков при цитировании, прямой речи.

4) "Окрестность" Для увеличения количества сопоставленных предложений используется приближенный метод сопоставления предложений в некоторой заданной окрестности ε. Например, если есть сопоставление $A(k)->B(n)$, $A(k+t)->B(n+t)$, где $0<t<\varepsilon$, то сопоставляем и $A(k+i)->B(n+i)$ для всех $0<i<t$. По умолчанию используется окрестность, состоящая из одного предложения, т.е. $\varepsilon=1$.

Кроме этого, стоит отметить, что алгоритмы для решения данной задачи можно использовать не только в последовательных рекурсивных вызовах, но и совместно.

Для каждого алгоритма определим функцию *возможного сопоставления предложений AlgorithmSentenceCorresponding(x, y)*, где $x \in \{1, ..., n\}$, $y \in \{1, ..., m\}$. Область значений – действительные числа. Данная функция сопоставляет каждой возможной паре, состоящей из предложения и перевода, некоторое число, которое характеризует степень их соответствия. По умолчанию значение равно нулю, что соответствует тому случаю, когда данный алгоритм не в состоянии оценить соответствие. В зависимости от возможности сопоставить предложение и перевод в данной паре,

функция имеет положительное или отрицательное значение, что соответствует случаям, когда предложения скорее сопоставимы или скорее не сопоставимы соответственно. Чем больше по модулю значение функции, тем сильнее утверждение.

Когда известны значения функций от каждого алгоритма, нужно выбрать пары, которые будут сопоставлены на данной итерации. Для этого можно вычислить среднее арифметическое значений всех функций для каждой пары и выбрать из них те, которые выше некоторого заданного порога. Кроме этого, можно принять во внимание солидарность или противоречивость мнений алгоритмов о сопоставлении той или иной пары (т.е. принять во внимание знаки соответствующих значений функций).

### *Пополнение словаря*

Алгоритм №1: Пополнение англо-русского словаря

Вход: множество пар предложений из текстов и переводов.

Выход: созданный (пополненный) словарь для данной предметной области.

Примерная схема работы алгоритма. Рассматриваем отдельно каждую сопоставленную пару. В предложении и его переводе для каждого слова или имеющегося в словаре словосочетания (кроме предлогов, служебных слов и других слов из специального словаря исключений) находим его начальную форму, используя морфологический модуль для соответствующего языка. Затем для слов из предложений на английском языке проверяем, есть ли их начальные формы в нашем базовом словаре и словаре предметной области. Таким образом, получаем для каждой пары предложение-перевод множество слов на английском языке, которых нет в словаре. Слова из предложений на русском языке мы рассматриваем все полностью.

Например, для *i*-й пары имеем множество новых английских слов $eWordSet(i)=\{eWord(i, 1), \ldots, eWord(i, n(i))\}$ и множество всех русских слов $rWordSet(i)=\{rWord(i, 1), \ldots, rWord(i, m(i))\}$.

Рассмотрим множества всех новых английских и всех русских слов из всех пар: $eWordSet = eWordSet(1) \cup \ldots \cup eWordSet(n)$ и $rWordSet = rWordSet(1) \cup \ldots \cup rWordSet(m)$.

Теперь нам необходимо корректно сопоставить слова из множеств *eWordSet* и *rWordSet,* чтобы добавить их в словарь.

Зададим *функцию возможного сопоставления слов* следующим образом:

$$wordCorrespondence(eWord, rWord) = \frac{|\{i : eWord \in eWordSet(i) \wedge rWord \in rWordSet(i)\}|}{|\{i : eWord \in eWordSet(i)\}|}$$

Последовательно фиксируем каждое слово $eWordSelected \in eWordSet$. Для него находим максимум функции *wordCorresponding(eWordSelected, rWord)* по словам $rWord \in rWordSet$ таких частей речи, чтобы было возможно, что *rWord* – перевод *eWordSelected.* Если русское слово *rWordSelected*, на котором достигается этот максимум, единственно, то считаем, что слова *eWordSelected* и *rWordSelected* нужно сопоставить и добавить эту пару в словарь. Другими словами, для каждого английского слова мы находим такое русское, с которым они наиболее часто встречались в парных множествах новых слов, полученных из предложений. Если таких слов несколько, то приоритет отдается тем, которых нет в словаре. Если и после этого осталось более одной альтернативы, то предлагается выбрать подходящее слово из имеющихся слов вручную.

Полученный список сопоставленных слов может быть предложен человеку на утверждение. Если какому-то английскому слову русское слово сопоставлено некорректно, то имеется возможность посмотреть другие русские слова с достаточно большими значениями функции, либо вообще все возможные слова из соответствующих предложений.

В некоторых случаях русское слово есть транслитерация английского слова. Это тоже можно с успехом использовать при сопоставлении.

К достоинствам данного алгоритма можно отнести относительное быстродействие и простоту реализации в сравнении со следующим алгоритмом. Однако у этого алгоритма есть недостатки. В случае, когда в словарь должен добавиться новый смысл уже существующего английского слова, существующее слово не будет включено в список рассматриваемых слов, сопоставление придется делать вручную. Кроме этого в данном алгоритме не производится выделение словосочетаний.

Алгоритм №2: Пополнение семантических сетей WordNet и RussianWordNet

Рассмотрим более сложный алгоритм, использующий семантические сети WordNet и RussianWordNet.

Вход: множество пар предложений из текстов и переводов.

Выход: созданный (пополненный) словарь для данной предметной области или созданные (пополненные) семантические сети WordNet и RussianWordNet.

Примерная схема работы алгоритма. Рассматриваем отдельно каждую сопоставленную пару. В предложении и его переводе для каждого слова или имеющегося в словаре словосочетания (кроме предлогов, служебных слов и других слов из специального словаря исключений) находим его

начальную форму, используя морфологический модуль для соответствующего языка.

Пусть *eWord(1), …, eWord(n)* и *rWord(1), …, rWord(m)* – начальные формы слов из предложения и его перевода соответственно. Для каждого из них получаем множество синсетов, в которые они входят (для первого из WordNet: *eSynsets(1), …, eSynsets(n);* для второго из RussianWordNet: *rSynsets(1), …, rSynsets(m)*). Далее находим все эквивалентные связи между найденными синсетами, когда английский и русский синсеты выражают одну и ту же смысловую единицу. Те слова, для синсетов которых найдены подобные связи, далее не рассматриваем. Оставшиеся без связей слова предлагаем сопоставить, возможно, создавая новые синсеты.

### *Итерационный характер работы алгоритмов*

Результатом работы алгоритмов является построенный (пополненный) словарь для заданной предметной области. Заметим, что на этапе 2 построения словаря (Сопоставление предложений и переводов) в алгоритме 2 (Однозначно переводимые слова) используется уже имеющийся словарь для предметной области. Таким образом, мы можем организовать итерационный процесс построения словаря. Будем повторять последовательно этапы 2 и 3 до тех пор, пока словарь не перестанет пополняться.

### *Реализация алгоритмов*

Были полностью реализованы алгоритмы для первой и второй подзадач. Для третьей подзадачи был реализован алгоритм, основанный на англо-русском словаре. Проверка более сложного алгоритма, использующего семантические сети WordNet и RussianWordNet, проводилась вручную.

Алгоритмы были реализованы на языке программирования Visual C#. В качестве англо-русского словаря с общей лексикой использовался словарь Мюллера. Использовался морфологический модуль проекта «Автоматическая Обработка Текста» (http://www.aot.ru/). При разработке данного модуля для реализации английской морфологии использовался WordNet, для русской морфологии – словарь Зализняка.

### *Анализ результатов работы*

Проводилось тестирование программной системы по текстам различных тематик: информационные технологии, биология, математика, химия, физика и др. В настоящее время с помощью описанных алгоритмов распознается корректно более 95% предложений – первая подзадача, сопоставляется в среднем 60% предложений и переводов – вторая подзадача. С помощью простого алгоритма сопоставления находится в результате до 65% терминов из текстов – третья подзадача. Никакого противоречия в том, что процент найденных терминов больше, чем процент сопоставленных предложений, нет, т.к. в большинстве случаев число новых терминов в тексте гораздо меньше общего числа предложений. Сопоставляются правильно от 65% до 95% слов, часть из них сопоставляется автоматически. Вообще процент слов, сопоставляющихся автоматически, очень сильно зависит от типа текста (научная статья, тезисы, публикация в СМИ и т.д.) и стиля изложения, поэтому здесь нельзя привести никакие усредненные значения.

### *Заключение*

При выполнении работы исследована возможность автоматизации процесса построения (пополнения) словарей для заданной предметной области. Проведен анализ данной задачи, позволивший разбить ее на более мелкие подзадачи. Для решения каждой подзадачи разработаны алгоритмы. Для большинства алгоритмов выполнена программная реализация. Проведен анализ и сравнительная оценка полученных результатов по различным предметным областям.

В результате был сделан вывод, что можно значительно автоматизировать процесс построения и пополнения специализированных словарей.

### *Дальнейшая работа*

Работы в данном направлении могут быть продолжены. Остались не решенными некоторые возникшие задачи. В настоящее время планируется работа по описанным ниже направлениям.

Проблема автоматического выделения словосочетаний. В настоящее время разработанные алгоритмы не поддерживают автоматическое добавление в словарь терминов, состоящих из нескольких слов. Такие термины нужно добавлять в ручном режиме. Требуется автоматизировать алгоритм нахождения словосочетаний и отделения их от терминов, состоящих из одного слова.

Разработка специальных модулей для учета разметки документов. В настоящее время в сети Интернет многие тексты представлены не в простом текстовом формате, а в форматах с внутренней разметкой. К таким форматам можно отнести HTML, XML, LaTeX и другие. При решении задачи выравнивания текстов, находящихся в данных форматах, мы можем использовать кроме собственно самого текста информацию о разметке документов. Требуется написать программные модули для поддержки различных форматов разметки.

***Литература***


1. Гельфейнбейн И.Г., Гончарук А.В., Лехельт В.П., Липатов А.А., Шило В.В. Автоматический перевод семантической сети WORDNET на русский язык // «Компьютерная лингвистика и интеллектуальные технологии» Труды Международного семинара Диалог'2003. М.: Наука, 2003, С. 193-198.
2. Lipatov A, Goncharuk A, Helfenbein I, Shilo V, Lehelt V Automatic Creation of Non-English WordNet-like Lexical Databases // Lecture Notes in Computer Science, Japan: Papillon Workshop 2003.
3. Miller G. Five Papers on WordNet. // Technical Report, USA: Princeton University, Cognitive Science Laboratory,.
4. Azarova I, Mitrofanova O, Sinopalnikova A, Yavorskaya M, Oparin I. RussNet: Building a Lexical Database for the Russian Language.
5. Азарова И.В., Митрофанова О.А., Синопальникова А.А., Ушакова А.А., Яворская М.В. Разработка компьютерного тезауруса русского языка типа WordNet.
6. Липатов А.А., Шило В.В. Автоматизация процесса создания и пополнения специализированных словарей // Сборник тезисов XLII Международной студенческой конференции «Студент и научно-технический прогресс», Новосибирск.
7. Лехельт В.П. Усовершенствованный Автоматический перевод семантической сети WORDNET на русский язык, дипломная работа.
8. Каневский Е.А. Некоторые вопросы пополнения морфологического словаря терминами предметной области // труды конференции «Диалог-2001», М.: Наука
9. Игитов С.В., Крючкова Е.Н., Старцев П.Л., Чумак М.В. Проблемы формирования словарей в системах искусственного интеллекта.
10. Поминов А.В. Некоторые вопросы построения многоязычных автоматических словарей.
11. Лукашевич Н.В., Добров Б.В. Двуязычный информационный поиск на основе автоматического концептуального индексирования // «Компьютерная лингвистика и интеллектуальные технологии» Труды Международного семинара Диалог'2003. М.: Наука, 2003, С 425-429