Technology of the automated thesaurus construction for Information Retrieval

Alexander Panchenko
Bauman Moscow State Technical University
panchenko.alexander@gmail.com

Abstract

This paper presents a technology of the automated construction of a thesaurus of a domain based on processing a corresponding text corpus of the domain. The developed technology assists expert on each step of the construction of a semantic resource by calculating sets of word-, phrase-, and relation-candidates for including into a thesaurus.

Технология автоматизированного построения информационно-поискового тезауруса

Александр Панченко
МГТУ им. Н.Э.Баумана
panchenko.alexander@gmail.com

1. Введение

Информационно-поисковый тезаурус – мощный инструмент для организации, индексирования и управления большими объемами данных. Главные цели использования тезауруса – предоставление стандартного набора лексики для индексирования документов и иерархий семантических связей между концептами тезауруса для переформулирования поисковых запросов и некоторых других задач [2, 3].

Одним из барьеров для широкого использования тезаурусо-ориентированных подходов в информационных системах является высокая сложность и стоимость ручного создания семантического ресурса.

В работе было проведено исследование методов и алгоритмов, которые могли бы помочь автоматизировать составление такого словаря; была изучена технология и выявлены основные этапы процесса ручного составления тезауруса, после чего было предложено автоматизировать процесс посредством создания информационной человеко-машинной

технологии построения тезауруса. Разработка технологии является ключевой частью работы и основывается на комбинации машинных методов анализа корпуса текстов предметной области и ручной работы эксперта.

2. Информационно-поисковые тезаурусы

Информационно-поисковый тезаурус представляет собой составленный по определенным правилам словарь терминов и словосочетаний, для определенной предметной области, создаваемый для улучшения качества информационного поиска в данной предметной области.

Два основных применения информационно-поисковых тезаурусов это индексирование документов с использованием *концептов* содержащихся в семантическом ресурсе и использование иерархических, ассоциативных и синонимичных связей при обработке поисковых запросов пользователя. Помимо этого, семантические связи между дескрипторами могут быть использованы для классификации и рубрикации документов, составлении списка связанных с запросом слов и некоторых других задачах информационного поиска. На следующем рисунке изображен пример использования тезауруса в информационно-поисковой системе:

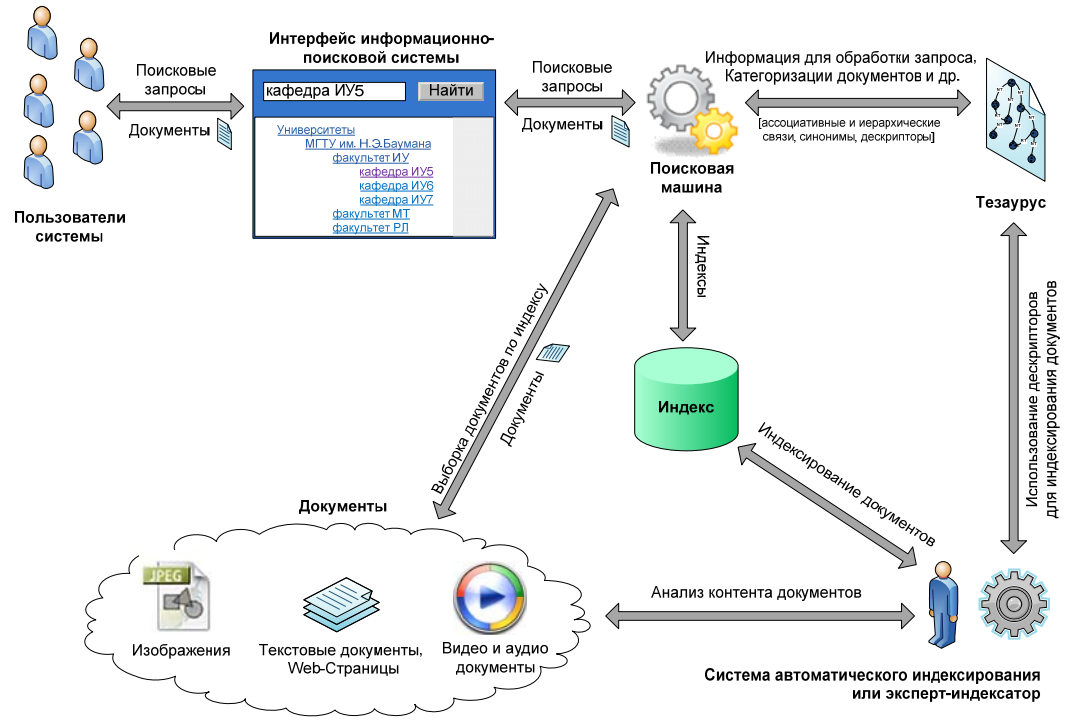

Рисунок 1. Пример использования тезауруса в информационно-поисковой системе.

Тезаурус состоит из *концептов*, которые обычно содержащих один или несколько дескрипторов связанных отношением синонимии или квази-синонимии. *Дескрипторы*

представляют собой слова и словосочетания, отражающие основные понятия предметной области. При этом, концепты могут быть связаны ассоциативными, иерархическими и некоторыми другими типами отношений.

Примером концепта может служить набор дескрипторов: “охрана природы”, “защита природы”, “природоохранная сфера”, “природоохранительный”, “природоохранный” относящиеся к одному понятию. На следующем рисунке приведен концепт “средства транспорта” (means of transport) европейского тезауруса Eurovoc.

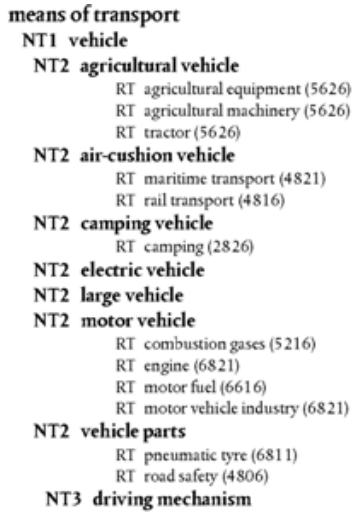

Рисунок 2. Концепт “средства транспорта” тезауруса Eurovoc.

Информационно-поисковый тезаурус – словарь, составляемый вручную экспертом-лингвистом, специалистом в области построения словарей и семантических ресурсов. При составлении подобного словаря стоит задача получить тезаурусное описание одной или нескольких предметных областей, при этом, часто существует корпус текстов который является основой для создания словаря. Эксперт проводит анализ корпуса текстов и руководствуясь технологией ручного построения тезауруса [4], описанной в стандартах [5] составляет список терминов описывающих заданную предметную область и включает их тезаурус в качестве дескрипторов. После этого термины группируются в концепты и между ними устанавливаются иерархические и ассоциативные отношения.

Для процесса ручного создания тезауруса характерны такие недостатки как высокая стоимость и длительность создания ресурса, обусловленность результата от квалификации эксперта, невозможность вручную проанализировать весь корпус текстов и некоторые другие. Было предложено автоматизировать процесс создания тезауруса с помощью технологии автоматизированного построения тезауруса, которая реализуется в Автоматизированной Системе Построения Тезауруса (см.рисунок 3).

В результате исследования был выявлен круг существующих программных решений для работы с информационно-поисковыми тезаурусами: Oracle Text / Thesaurus Management

System, Amicus Thesaurus, Livelink Collections Server Thesaurus Manager, Thesaurus Master и некоторые другие.

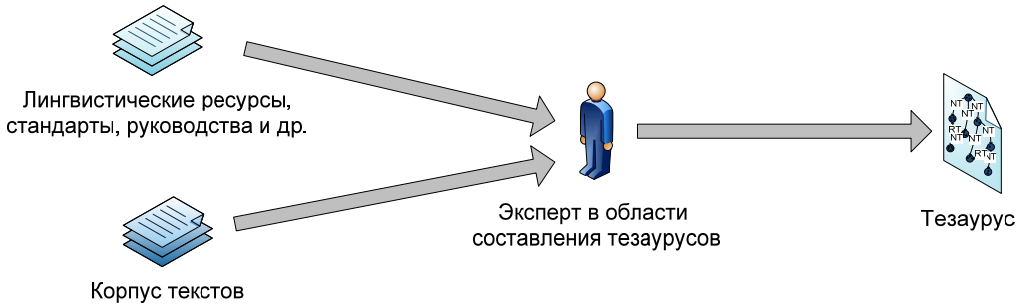

Рисунок 3.1 Процесс ручного построение тезауруса.

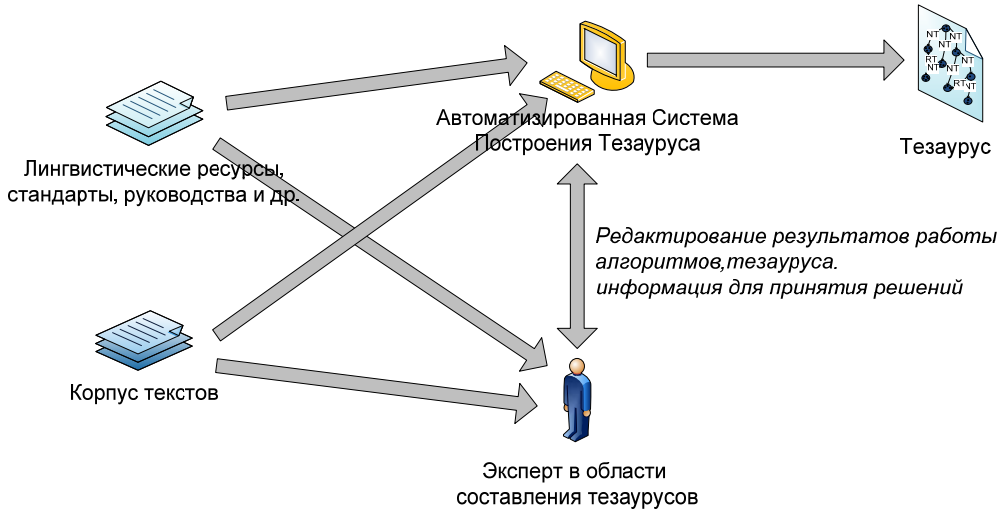

Рисунок 3.2 Процесс автоматизированного построение тезауруса.

Оказалось, что ни один из найденных продуктов не поддерживает процесс автоматизированного построения тезауруса – программные решения лишь предоставляют развитые возможности для хранения и управления заранее созданными вручную контролируемыми терминологиями.

3. Технология автоматизированного построения тезауруса

Попытки создать тезаурус автоматически уже предпринимались, так Филлипович [7] предложил технику построения тезауруса текстов на основе статистики Вилкоксона, Грегори [11] описывает ряд методов для автоматического построения тезауруса, в частности подобные используемым в части 7 данной статьи. Можно также отметить работы посвященные автоматизации отдельных этапов построения тезауруса, к примеру, извлечения терминологий из корпуса текстов [12]. В данной статье предлагается новая технология автоматизированного построения тезауруса, включающая в себя описание не

только алгоритмической части, но и задания характера разделения работы между человеком и системой при составлении словаря.

Представляемая технология базируется на так называемом *дедуктивном* подходе к составлению тезауруса, который подразумевает, что все дескрипторы тезауруса извлекаются из корпуса текстов. При ручном составлении эксперт нередко добавляет термины не присутствующие в корпусе, однако количество подобных слов и словосочетаний, как правило, не превышает 20-25%. Особенности предлагаемой технологии позволяют эксперту добавить недостающие термины вручную, при необходимости.

Главное отличие от ручного составления заключается в том что производится предварительная машинная обработка корпуса текстов и эксперту на каждом из этапов конструирования тезауруса представляются множество слов, словосочетаний и связей между ними – кандидатами для включения в тезаурус. Согласно технологии человек всегда сам определяет окончательный список элементов словаря, имея возможность добавления, удаления и модифицирования списка кандидатов, предложенных системой. Технология автоматизированного построения тезауруса состоит из нескольких этапов:

1) Предварительная обработка корпуса текстов: выполняются простейшие, но необходимые преобразования коллекции документов с целью представления их в виде пригодном для дальнейшей обработки с помощью методов машинного анализа текстовой информации.
2) Построение множества предпочтительных дескрипторов. Формирование множества слов и словосочетаний кандидатов для включения в тезаурус. Эксперт руководствуясь множеством кандидатов составляет список ключевых понятий предметной области.
3) Поиск в словаре синонимов отношений связей между дескрипторами тезауруса, редактирование экспертом автоматически найденных отношений и окончательное группирование дескрипторов в концепты.
4) Построение множества ассоциативных и иерархических отношений. Формирование множества не типизированных отношений кандидатов между дескрипторами тезауруса. Эксперт, используя построенное множество, определяет окончательное множество иерархических и ассоциативных связей между концептами тезауруса.

Кроме указанных четырех этапов в процесс обработки может быть осуществлен еще один – кластеризация документов корпуса (обозначен буквой N на рисунке 3), который требуется, если корпус текстов плохо структурирован и содержит документы, относящиеся к разным предметным областям. Данный этап подробно не рассматривается в статье.

В четвертой части статьи будет более подробно описан этап предварительной обработки корпуса текстов. О деталях процесса отбора слов и словосочетаний для словаря рассказано в пятой части статьи. Шестая и седьмая части раскрывают детали поиска отношений синонимии и иерархических отношений соответственно. На рисунке 4 представлена обзорная схема, показывающая основные этапы технологии автоматизированного построения тезауруса.

4. Предварительная обработка корпуса текстов

Большой корпус текстов может исчисляться в десятках, сотнях тысяч, а иногда и миллионах документов. Количество слов в таких корпусах текстов нередко достигает десятков миллионов. При этом, текст написанный на естественном языке слабо структурирован и может содержать большое количество ошибок.

Выделение значимых слов и словосочетаний в таком корпусе и применение алгоритмов анализа текстов требует предварительной обработки корпуса. В описываемой технологии были выбраны следующие этапы предварительной обработки текста:

4.1 Извлечение всех слов и словосочетаний из корпуса текстов

Множество всех уникальных слов содержащихся в корпусе TOK формируется с помощью процедуры *токенизации.* Алгоритм определят границы слов с помощью множества стоп-знаков слова – множества знаков, которые позволяют отделять в тексте слова друг от друга [7], а также некоторых правил, после чего составляется список всех уникальных слов в корпусе (иногда лемматизированных – см. следующий параграф). Размер множества TOK, как правило, не превышает 50000 слов. Будем считать также что TOK содержит пустую строку.

Важно извлечь из текста не только отдельные слова, но и словосочетания т.к. основные понятия предметной области очень часто представлены составными словами, к примеру, словосочетание “сельское хозяйство” с гораздо большей вероятностью будет включено сельскохозяйственный тезаурус в качестве дескриптора "сельское хозяйство", нежели слова "сельское" и "хозяйство" по отдельности. По соображениям целесообразности, мы ограничиваем максимальную длину искомых дескрипторов пятью словами. При этом, отдельный дескриптор d можно представить как 5-компонентный упорядоченный кортеж $\langle w_1, w_2, w_3, w_4, w_5 \rangle \in \mathrm{TOK}^5$. Можно получить представление множества всех отдельных слов TOK как множество кортежей $W \in \mathrm{TOK}^5$ если каждому слову $w \in TOK$ поставить в соответствие кортеж вида $\langle w, 0,0,0,0 \rangle$, где $0 -$ пустая строка.

На данном этапе формируется множество всех значимых словосочетаний в корпусе $MWE \in \mathrm{TOK}^5$, длиной от двух до пяти слов. Каждое такое словосочетание

представляется в виде кортежа вида $\langle w_1, w_2, w_3, w_4, w_5 \rangle$, при этом, кортеж соответствующий словосочетанию из трех слов будет иметь вид $\langle w_1, w_2, w_3, 0,0 \rangle$.

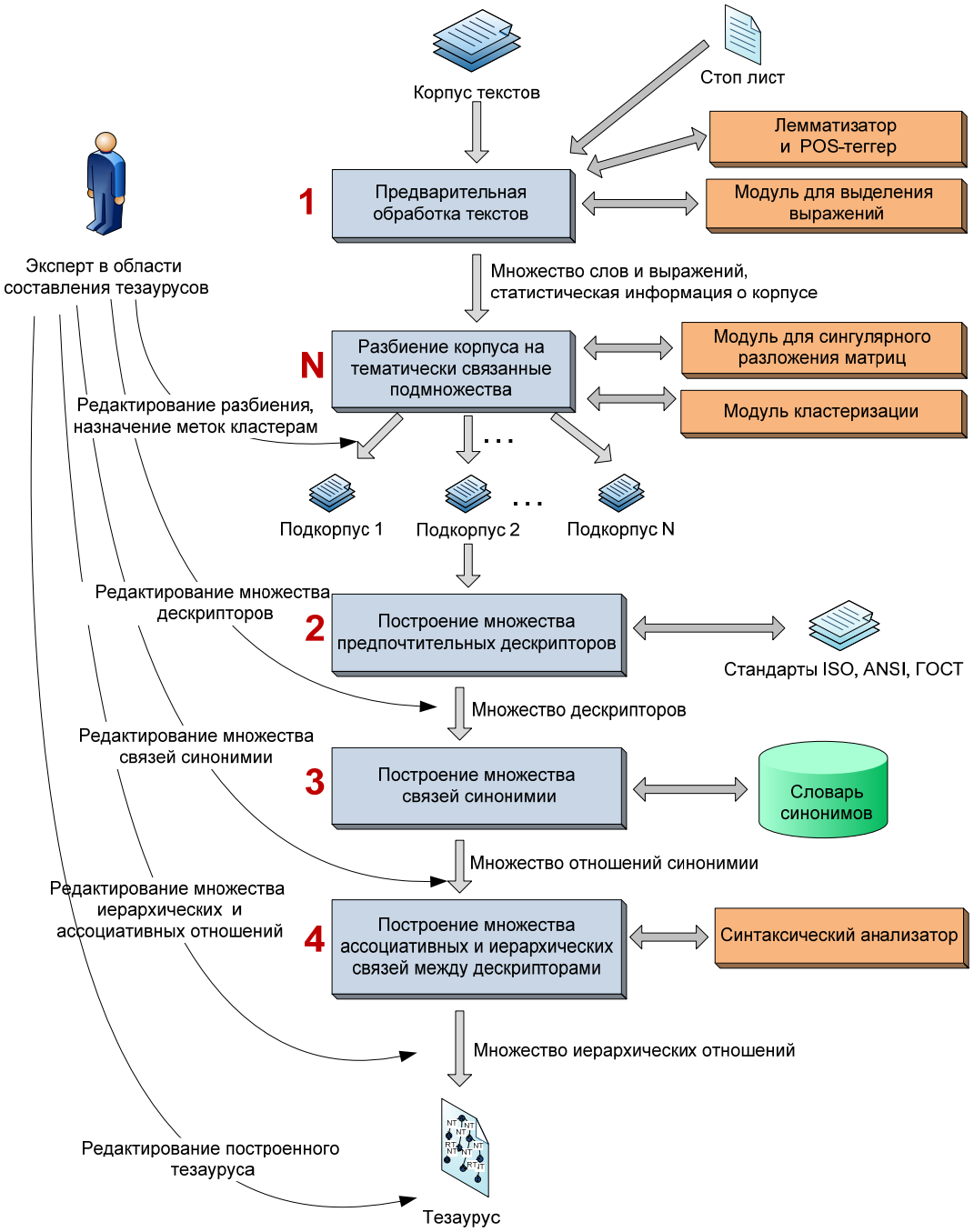

Рисунок 4. Технология автоматизированного построения тезауруса.

Для отыскания множества значимых словосочетаний были применены конечные преобразователи (finite state transducers), использующие специально подготовленные словари терминов, из которых может быть составлено словосочетание [1]. В результате данного этапа формируется множество дескрипторов корпуса $D = W \cup MWE$.

Для целей дальнейшего анализа важно не только определить множество слов и словосочетаний содержащихся в корпусе, но и сохранить информацию о том в каком предложении какой дескриптор содержится. Пусть в обрабатываемом корпусе содержится k

предложений, тогда будем ассоциировать число из последовательности $SNT = \{1,2,\dots,k\}$ с предложением. Определим соответствие $snt\_d \subseteq SNT \times D$ из множества предложений корпуса SNT во множество из дескрипторов корпуса D посредством задания множества всех упорядоченных пар $(x, d), x \in SNT, d \in D$.

4.2 Лемматизация слов

Лемматизация извлеченных слов – получение канонической формы слова, т.е. сведение словоизменительных форм слова к исходной (для существительных это именительный падеж, единственное число; для прилагательных это именительный падеж, единственное число, мужской род и т.п.). В лемматизированном тексте можно связать одинаковые слова текста находящиеся в разных словоформах.

4.3 Установление частей речи

Установление частей речи извлеченных слов и выражений производится с помощью специального модуля в синтаксическом анализаторе, который назначает каждому слову $d_i \in W$ часть речи pos_{d_i}. Данный этап делает возможным фильтрацию слов и словосочетаний по части речи, так, к примеру, из слов кандидатов в дескрипторы исключаются глаголы и глагольные конструкции, а предпочтение отдается существительным.

4.4 Синтаксический анализ корпуса

Синтаксический анализ – процесс определения грамматической структуры предложений составляющих текст. Результатом такого анализа является множество синтаксических зависимостей между лексическими единицами предложения, которые могут быть представлены тройками вида $\langle w_1, type, w_2 \rangle$, где $type -$ тип зависимости, а $w_1,\ w_2 -$ слова. Результатом работы синтаксического анализатора является дерево синтаксических зависимостей между лексическими элементами предложения для каждого из предложений текста. Множество разных синтаксических зависимостей, полученных в результате анализа обозначим как $SR_{corpus} = \{sr_1, \dots, sr_z\} \subseteq SR = TOK \times T \times TOK$, где i-е синтаксическое отношение sr_i представляет собой упорядоченную тройку $\langle w_{i1}, type_i, w_{i2} \rangle, w_{i1},\ w_{i2} \in TOK,\ type_i \in T -$ тип синтаксической зависимости. Ниже приведем пример результаты работы синтаксического анализатора предложения *«Лемматизация играет важную роль в задачах компьютерной лингвистики»:*

$$sr_1 = \langle \text{"лемматизация"}, \text{подлеж}, \text{"играть"} \rangle$$

$$sr_2 = \langle \text{"играть"}, \text{сказ}, \text{"лемматизация"} \rangle$$

$$sr_3 = \langle \text{"играть"}, \text{прям\_дополн}, \text{"роль"} \rangle \dots$$

Множество T зависит от используемого анализатора и может содержать большое

количество разнообразных отношений, однако в используемой модели были использованы восемь типов отношений, формирующие множество $T_{sel} \subseteq T$, перечисленных в таблице 1.

Таблица 1. Типы используемых синтаксических зависимостей.

	Описание отношения	Формальное обозначение*
1	w – подлежащее глагола v	$\langle w, \text{подлеж}, v\rangle$
2	w – прямое дополнение глагола v	$\langle w, \text{прям\_дополн}, v+p\rangle$
3	w – предложное дополнение глагола v введенного предлогом p	$\langle w, \text{пред\_доп}, v+p\rangle$
4	w – наречие глагола v введенное предлогом p	$\langle w, \text{нарч}, v\rangle$
5	w модифицировано прилагательным a	$\langle w, \text{прил}, v\rangle$
6	w относится к приложению n	$\langle w, \text{прилож}, n\rangle$
7	w – сказуемое существительного n	$\langle w, \text{сказ}, n\rangle$
8	w относится к сочинению n	$\langle W, \text{сочин}, n\rangle$

*прим. $w, v, p, n \in TOK$.

Из всего множества извлеченных отношений зададим подмножество $SR_{sel} \subseteq SR$ отношений все элементы которого имеют один из восьми перечисленных типов синтаксической связи:

$$SR_{sel} = \{sr_i : sr_i \in SR_{corpus} \wedge type_i \in T_{sel}\}.$$

Для целей дальнейшего анализа важно сохранить информацию о том в какой набор синтаксических зависимостей был получен для каждого из предложений в результате анализа. Для этого определим соответствие $snt\_sr \subseteq SNT \times SR_{sel}$ из множества предложений корпуса SNT во множество зависимостей между лексическими единицами корпуса SR_{sel}, посредством задания множества всех упорядоченных пар $(x, sr), x \in SNT, sr \in SR_{sel}$ соответствующих результатам разбора корпуса текство анализатором.

4.5 Декапитализация, деакцентизация и удаление стоп-слов

Декапитализация – преобразование всех символов корпуса к нижнему регистру. *Деакцентизация* – прием используемый при обработке текстов на французском языке, в которых существуют буквы с акцентами, к примеру é, è, à и т.п. Из-за особенностей грамматики, одно и то же слово в разных контекстах появляться с акцентами и без, поэтому все символы с акцентами заменяются на аналоги без акцентов. *Стоп лист* – список вспомогательные слов несущих мало информации о содержании документа, таких как артикли, союзы и наиболее распространенные глаголы. Из текста удаляются все слова из стоп листа.

4.6 Вычисление некоторых статистик для дескрипторов тезауруса

На данном этапе вычисляется частота встречаемости $\#d_i$ и коэффициент $tf - idf$

для каждого из l найденных дескрипторов $d_i \in D$:

$$tf - idf(d_i) = \frac{\#d_i}{\sum_{j=1}^{l} \#d_j} \cdot \log \frac{|DOC|}{|\{doc: d_i \in doc\}|}, \qquad (1)$$

где $|DOC|$ — количество документов в корпусе, а $|\{doc: d_i \in doc\}|$ — количество документов, в которых встречается дескриптор d_i. Кроме этого, на данном этапе формируется множество $H \subseteq D$ слов встречающихся в заголовках документов, такое что $H = h_1 \cup h_2 \cup \ldots \cup h_{|DOC|}$, $h_i = \{d: d$ находится в заголовке $doc_i\}$.

5. Построение множества дескрипторов тезауруса

Для того чтобы выделить из всех извлеченных дескрипторов ключевые понятия предметной области строиться сюрьективное отображение с помощью функции отбора дескрипторов $f_{sel}: D \to A$, где $A -$ множество действительных весов a_i лежащих в интервале [0;1]. Здесь каждый дескриптор d_i представляется следующим вектором координат $\boldsymbol{d_i} = (d_i, \#d_i, tf - idf_{d_i}, mwe_{d_i}, head_{d_i}, pos_{d_i})$, где $tf - idf_{d_i} = tf - idf(d_i)$,

$$mwe_{d_i} = mwe(d_i) = \begin{cases} 0, & d_i \in W \\ 1, & d_i \in MWE \end{cases}, \qquad head_{d_i} = head(d_i) = \begin{cases} 1, & d_i \in H \\ 0, & \text{иначе} \end{cases}.$$

Таким образом, вес дескриптора d_i равен

$$a_i = f_{sel}(\boldsymbol{d_i}) = f_{sel}(d_i, \#d_i, tf - idf_{d_i}, mwe_{d_i}, head_{d_i}, pos_{d_i}). \qquad (2)$$

На основе рассчитанных весов строится множество дескрипторов-кандидатов для включения в тезаурус в которое включаются из дескрипторы d_i вес которых не менее заданного порога a_{min}:

$$D_{thes} = \{d_i: f_{sel}(\boldsymbol{d_i}) \geq a_{min}\}. \qquad (3)$$

Заключительной стадией данного этапа является анализ и ручное редактирование множества слов и словосочетаний кандидатов экспертом-составителем семантического ресурса (см. рисунок 1).

6. Построение множества отношений синонимии

Алгоритм построения множества отношений синонимии S_{thes} между элементами множества дескрипторов тезауруса D_{thes} основывается на отыскании отношений в словаре синонимов. Пусть S – множество всех связей синонимии содержащихся в базе синонимов, такое что $S = \{s_1, s_2, \ldots, s_m\}$, где s_i – отношение синонимии которое можно представить в виде неупорядоченной пары вида $\langle d_i, d_j \rangle$. Ниже приведен алгоритм построения множества отношений синонимии $S_{thes} \subseteq S$; множество S в алгоритме смоделировано как хэш-таблица содержащая в качестве значений хэш-таблицу с строками синонимов того или иного слова. Множество дескрипторов тезауруса D_{thes} смоделировано как хэш-таблица хранящая

множество строк с дескрипторами тезауруса (в данной реализации каждому кортежу $d \in D_{thes}$ соответствует строка полученная конкатенацией всех его компонент):

FIND SYN RELATIONS(S, D_{thes})

1 $S_{thes} \leftarrow \emptyset$
2 ***foreach*** $d \in D_{thes}$
3 $dIndex \leftarrow$ SEARCH(S, d)
4 ***if*** $dIndex$
5 ***foreach*** $syn \in S[dIndex]$
6 $synIndex \leftarrow$ SEARCH(D_{thes}, syn)
7 ***if*** $synIndex$
8 $S_{thes} \leftarrow S_{thes} \cup \langle d, syn \rangle$
9 ***return*** S_{thes}
10 ▷ функция SEARCH$(hashtable, string)$ возвращает индекс строки в таблице,
11 ▷ либо 0 если строка не найдена

Среднее время, требуемое для выполнения алгоритма, равно $O(|D_{thes}| \cdot L_{avg})$, где L_{avg} – среднее количество синонимов для слова из словаря синонимов; максимальное время выполнения алгоритма равно $O(|D_{thes}| \cdot L_{max})$, где L_{max} – максимальное количество синонимов для слова из словаря синонимов. После завершения работы алгоритма, согласно разработанной технологии, эксперт проводит анализ и редактирование множества S_{thes} вручную.

7. Построение множества иерархических и ассоциативных отношений

Как и для предыдущего этапа, задачу определения иерархических и ассоциативных отношений между дескрипторами тезауруса (элементами множества D_{thes}) можно разделить на 2 этапа: автоматическое построение множества отношений-кандидатов и ручное редактирование полученного множества отношений экспертом. Важным отличием от нахождения связей синонимии является то что поиск отношений производиться не в заранее составленной базе, а непосредственно в тексте, на основе анализа совместной встречаемости слов.

Задача первого этапа – построение множества отношений-кандидатов между дескрипторами тезауруса $R_{thes} = \{r_1, \dots, r_l\},\ r_i = \langle d_1, d_2 \rangle,\ d_i \in D_{thes}$. Ипользуемый на этом этапе метод основан на алгоритме анализа контекстов из синтаксических зависимостей между лексическими единицами текста [6, 10], который использует модель векторной алгебры [8, 9] для вычисления семантической близости между дескрипторами. Построение

множества отношений-кандидатов R_{thes} между дескрипторами тезауруса D_{thes} включает в себя следующие шаги:

1) Синтаксический анализ корпуса текстов (см.раздел 4.4)
2) Построение множества признаков, определение базиса признаков
3) Вычисление координат дескрипторов в базисе признаков
4) Вычисление матрицы подобия дескрипторов тезауруса
5) Построение множества отношений между дескрипторами

На втором этапе полученное множество отношений редактируется экспертом: удаляются некорректные связи, добавляются недостающие. Кроме этого, составитель словаря должен определить тип отношения т.к. автоматическая процедура не типизирует найденные отношения.

7.1 Построение множества признаков, определение базиса признаков

Под *признаком* будем понимать упорядоченную пару $\beta_i = \langle type_i, w_i \rangle,\ type_i \in T, w_i \in TOK$. Множество всех возможных признаков задается декартовым произведением $T \times TOK$. Определим сюръективное соответствие $sr\_B: SR \to T \times TOK$, которое каждому синтаксическому отношению $\langle w_{i1}, type_i, w_{i2} \rangle$ ставит в соответствие признак равный $\langle type_i, w_{i2} \rangle$. В рассматриваемой модели для описания дескрипторов тезауруса используется *множество признаков* мощности m, которое обозначим как $B \subseteq T \times TOK$. Множество признаков включает в себя все различные признаки соответствующие множеству синтаксических зависимостей SR_{sel}. Множество B вычисляется с помощью алгоритма CALCB&C (см.7.2).

7.2 Вычисление координат дескрипторов в базисе признаков

В используемой модели каждый дескриптор тезауруса d_i представляется линейной комбинацией признаков, записанной в виде *характеристического вектора* $\boldsymbol{f_i}$=(f_{i1}, … f_{im}) в стандартном базисе $\boldsymbol{b} = (\boldsymbol{b_1} \ldots \boldsymbol{b_m})$, $b_i \in \mathbb{R}^m$:

$$\boldsymbol{b_1} = (1, 0, \ldots, 0),$$
$$\boldsymbol{b_2} = (0, 1, \ldots, 0),$$
$$\ldots$$
$$\boldsymbol{b_m} = (0, 0, \ldots, 1)$$

где каждому из базисных векторов $\boldsymbol{b_i}$ соответствует синтаксический признак β_i. Так как мощность множества признаков B по определению равна размерности базиса векторного пространства $\boldsymbol{b}$ и потому что все элементы множества B различны можно задать биективное отображение $B\_b: B \to \boldsymbol{b}$ которое ставит в соответствие каждому признаку β_i вектор базиса $\boldsymbol{b_i}$.

Обозначим как C^* мультимножество составленное из пар $\langle d_i, \beta_j \rangle$ которое задает связь

между дескрипоторами тезауруса и множеством признаков. Координаты вектора $\boldsymbol{f_i}$ в базисе $\boldsymbol{b}$, моделирующие дескриптор d_i будем определять как количество соответвующих синтаксических зависимостей относящихся к дескриптору которое равно кратности $\#\langle d_i, \beta_j \rangle$ соответствующих элементов мультимножества C^*. Таким образом, будем определять координаты следующим образом

$$f_{ij} = \#\langle d_i, \beta_j \rangle,$$

$$\boldsymbol{f_i} = \sum_{\beta_j \in B} \#\langle d_i, \beta_j \rangle B\_b(\beta_j) = \sum_{j=1}^{m} \#\langle d_i, \beta_j \rangle \boldsymbol{b_j}. \qquad (4)$$

Для вычисления множества признаков B и мультимножества C будем использовать следующий алгоритм:

CALC B&C (SR_{sel})[1]

1 $B \leftarrow \emptyset$

2 $C^* \leftarrow \emptyset$

3 ***for*** $i \leftarrow 0$ ***to*** k

4 ▷ Сохраняет дескрипторы тезаурус для обрабатываемого предложения

5 $D_i \leftarrow snt\_d(i)$

6 ***foreach*** $sr_i \in snt\_sr(i)$

7 ▷ Здесь подразумевается что $sr_i = \langle w_{i1}, type_i, w_{i2} \rangle$

8 $d = b(w_{i1}, D_i)$

9 ***if*** $d \neq \emptyset$

10 ▷ Синт. зависимость относиться к какому – либо дескриптору тезауруса

11 $C^* \leftarrow C^* \cup \langle d, sr\_B(sr_i) \rangle$

12 $B \leftarrow B \cup sr\_B(sr_i)$

13 ***return*** C^*

В алгоритме используется функция $b(w, D)$ для определения принадлежит ли слово множеству дескрипторов содержащихся в предложении

$$b(w, D_w) = \begin{cases} d, \text{если } (\exists d = \langle w, w_2, w_3, w_4, w_5 \rangle) \vee (\exists d = \langle w_1, w, w_3, w_4, w_5 \rangle) \vee (\exists d = \langle w_1, w_2, w, w_4, w_5 \rangle) \vee \\ \qquad (\exists d = \langle w_1, w_2, w_3, w, w_5 \rangle) \vee (\exists d = \langle w_1, w_2, w_3, w_4, w \rangle), \text{ где } d \in D_w, \ w_i \in TOK \\ \emptyset, \text{ иначе} \end{cases}.$$

Совокупность характеристических векторов, записанную в виде $(\boldsymbol{f_1}\ \boldsymbol{f_2}\ \ldots \boldsymbol{f_l})^{\mathrm{T}}$, будем называть *матрицей признаков* и обозначать как F. На рисунке 5 изображен пример матрицы признаков, полученной при обработке корпуса текстов сельскохозяйственной тематики.

7.3 Вычисление матрицы подобия дескрипторов тезауруса

Следующим этапом является определение степени семантической близости между

[1] В приведенном алгоритме вычисление множества B приведено для наглядности: как видно из (4), для вычисления координат достаточно установить множество C^*.

всеми дескрипторами тезаруса. Для этого совершается $\frac{l^2}{2} - l$ попарных сравнений между

	$\beta_1 = \langle$подлеж, работать$\rangle$	$\beta_2 = \langle$сказ, фермер$\rangle$	$\beta_3 = \langle$пр_дополн, трактор$\rangle$	…	$\beta_j = \langle type, d \rangle$	…	$\beta_m = \langle$подлеж, пахать$\rangle$	
$d_1 = \langle$промышленность, 0,0,0,0$\rangle$	*1*	*0*	*13*	…	…	…	*43*	f_1
$d_2 = \langle$сельское, хозяйство, 0,0,0$\rangle$	*23*	*1*	*98*	…	…	…	*32*	f_2
…	…	…	…	…	…	…	…	
$d_i = \langle w_1, w_2, w_3, w_4, w_5 \rangle$,	…	…	…	…	f_{ij}	…	…	f_i
…	…	…	…	…	…	…	…	
$d_n = \langle$фермер, 0,0,0,0$\rangle$	*43*	*32*	*45*	…	…	…	*1*	f_n

Рисунок 5. Матрица признаков.

векторами $\boldsymbol{f_1 f_2 \ldots f_l}$ с помощью определенной метрики определенной как $\cos(\boldsymbol{f_i}, \boldsymbol{f_j})$. Результаты попарного сравнения характерестических векторов можно записать в виде треугольной *матрицы подобия* M, такой что элемент на пересечании i-ой строки с j-м столбцом будет показвать меру близости между дескрипторами тезауруса d_i и d_j и вычисляется следующим образом

$$m_{ij} = \frac{\boldsymbol{f_i} \cdot \boldsymbol{f_j}}{|\boldsymbol{f_i}| \cdot |\boldsymbol{f_j}|} = \frac{\sum_{p=1}^{l} (f_{ip} \cdot f_{jp})}{\sqrt{\sum_{p=1}^{l} f_{ip}^2 \cdot \sum_{p=1}^{l} f_{jp}^2}}. \qquad (5)$$

7.4 Построение множества отношений между дескрипторами

На основании расчитанных мер связи между дескрипторами, производится формирование множества отношений между дескрипторами тезауруса R_{thes}. Во множество отношений включаются те отношения, $r_i = \langle d_1, d_2 \rangle$ сила связи которых s_{ij} не менее заданного порога m_{min} :

$$R_{thes} = \{\langle d_i, d_j \rangle : m_{ij} \geq m_{min}\}. \qquad (6)$$

Результатом данного этапа является множество отношений кандидатов R_{thes} между дескрипторами тезауруса без указания типа связи. Среди найденных отношений которые не являются ни ассоциативными ни иерархическими, кроме этого могут быть вычеслены и такие отношения которые вовсе являются ложными. Поэтому заключительным шагом является редактирование отношений кандидатов составителем лингвистичесого ресурса и задание для каждого и задание типа отношения.

Заключение

В работе представлена технология автоматизированного построения информационно-поискового тезауруса на основе корпуса текстов предметной области. В разработанной технологи производится предварительная машинная обработка корпуса текстов и эксперту на каждом из этапов конструирования тезауруса представляются множество слов, словосочетаний и связей между ними – кандидатами для включения в тезаурус.

Список использованных источников

[1]. Karttunen L. (2000). «Applications of Finite-State Transducers in Natural Language Processing». 5th International Conference on Implementation and Application of Automata. pp: 34 – 46.

[2]. Baeza-Yates R., Ribeiro-Neto B. (1999). «Modern Information Retrieval». Addison Wesley Longman Publishing Co. Inc. 163-173.

[3]. Frakes W., Baeza-Yates R.(1992), «Information Retrieval. Data Structures & Algorithms». Prentice Hall PTR; Facsimile edition. pp.: 161-197.

[4]. Aitchison. J.(2002) Thesaurus Construction and Use: A Practical Manual. Routledge, 4 edition.

[5]. American National Standards Institute. ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies, 2005.

[6]. Peirsman Y., Heylen K., Speelman D. (2007). «Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts» . In Proceedings of the CoSMO workshop, Roskilde, Denmark, pages 9-16

[7]. Филиппович Ю., Прохоров А.(2002) «Семантика информационных технологий: опыты словарно-тезаурусного описания». Изд-во МГУП, ISBN 5-8122-0367-9.

[8]. Berry M., Dumais S., O'Brien G. (1994). «Using Linear Algebra for Intelligent Information Retrieval». Society for Industrial and Applied Mathematics.

[9]. Berry M., Drmac Z.,Jessup R. (1999).Matrices, Vector Spaces, and Information Retrieval. Society for Industrial and Applied Mathematics.

[10]. Sahlgren M. (2006). «The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces». A PhD dissertation submitted to Stockholm University.

[11]. Gregory G. (1994). «Explorations in Automatic Thesaurus Discovery». The Springer International Series in Engineering and Computer Science.

[12]. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. (2003). «Формирование базы терминологических словосочетаний по текстам предметной области».