Формирование базы терминологических словосочетаний по текстам предметной области ♣

© Б.В.Добров
Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова; АНО Центр информационных исследований
dobroff@mail.cir.ru

Н.В.Лукашевич
Научно-исследовательский вычислительный центр МГУ им. М.В.Ломоносова; АНО Центр информационных исследований
louk@mail.cir.ru

С.В.Сыромятников
Факультет вычислительной математики и кибернетики МГУ им. М.В.Ломоносова; АНО Центр информационных исследований
syrom@hotbox.ru

Аннотация

Работа посвящена исследованию задач поддержки формирования и сопровождения больших информационно-поисковых тезаурусов. Рассматриваются разные методы автоматического формирования состава терминологии тезаурусов по текстам предметной области. Обсуждаются вопросы управления процессами выделения терминологических словосочетаний.

1. Введение

Терминологические (терминоподобные) словосочетания, которые ассоциированы с значимыми понятиями предметной области, играют большую роль в методах обработки и анализа текстов, служат основой для формирования информационно-поисковых тезаурусов, рубрикаторов, используются для визуализации результатов информационного поиска при интерактивном уточнении запроса [1, 16], известны применения для многоязычного поиска [13].

В данной работе рассматриваются задачи первоначального формирования состава информационно-поискового тезауруса по текстам предметной области, а также задачи пополнения уже существующего тезауруса.

Терминологические ресурсы представляют собой достаточно сложные структуры с многообразными связями между своими элементами. Отбор элементов и установление конкретного вида связи не должно [5, 6, 8] противоречить некоторому количеству принципов, которые кладутся в основу и определяют свойства терминологического ресурса. Эти ограничения носят операционный характер, что не позволяет их полностью алгоритмизировать.

При автоматическом выявлении терминологических словосочетаний мы можем оперировать только текстовыми формами, которые ищутся среди устойчивых словосочетаний, определяемых по корпусу текстов предметной области.

Границы понятия «устойчивое словосочетание» (англоязычный аналог «collocations») определены нечетко. Следуя [13], можно выделить два крайних случая. С одной стороны, *идиомы*, когда смысл словосочетания не может быть выведен из смысла составляющих его слов («*утечка мозгов*», «*корабль пустыни*»). С другой стороны, *свободные словосочетания*, смысл которых образуется объединением смыслов слов.

Проблема состоит в том, что вывод по значениям составляющих словосочетание слов при этом понимается в общеязыковом лингвистическом смысле, но словосочетание может обладать дополнительными экстралингвистическими отношениями, которые не всегда можно получить монотонно по иерархии значений составляющих слов. Словосочетание «*снос жилья*», полностью описываемое своими словами в лингвистическом смысле, имеет дополнительные отношения, связанные с «*аварийностью жилья*», «*переселением жильцов*» и т.п. При работе в предметной области возникают дополнительные сложности – свободное общеязыковое словосочетание «*большие окна*», может становиться термином в области строительства и торговли строительными товарами (попадая в фиксированные нормативные размеры).

В настоящей работе обсуждаются методы формирования базы терминологических словосочетаний, как описанные в литературе, так и методы, применяемые в практике ведения Общественно-политического тезауруса [5],

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

являющегося частью общего тезауруса русского языка РуТез [4], разрабатываемого АНО Центр информационных исследований. Тезаурус РуТез включает 44 тысячи понятий, 117 тысяч синонимов.

Общественно-политический тезаурус, насчитывающий 29 тысяч понятий и 75 тысяч синонимов, является основой лингвистической и тематической обработки полнотекстовых документов Университетской информационной системе РОССИЯ [9] (УИС РОССИЯ, www.cir.ru), поддерживаемой на базе НИВЦ МГУ.

2. Формирование начальной базы терминологических словосочетаний

В литературе [10-14] описано большое количество разнообразных методик первоначального формирования базы терминоподобных словосочетаний. Большинство методик имеет схожую структуру – в начале формируются списки «кандидатов», удовлетворяющих тем или иным лингвистическим условиям (синтаксическим ограничениям), затем производится сортировка полученного множества в соответствии с заложенной моделью образования терминов, с тем, чтобы предъявлять эксперту сначала словосочетания, имеющие больший вес.

Важнейшую роль играют различные дополнительные лексические фильтры, сокращающие выдачу за счет отбрасывания заведомо ложных вариантов, например, собственных имен персон, наименований географических объектов и т.п.

Естественный метод – сборка пар слов (возможно с предлогами), а затем упорядочивание их в соответствии с убывающей частотностью. Несмотря на простоту, данный метод формирует достаточно неплохую вершину списка при обработке больших объемов текстов.

Существует группа методов, где предложены другие критерии сортировки получающегося множества словосочетаний, отличные от простого подсчета суммарной частотности. Широко известен метод сортировки [10], учитывающий статистическую характеристику *«mutual information»* отношения вероятности совместной встречаемости двух слов в некотором текстовом окне к произведению вероятностей встречаемости каждого слова.

$$MI(a) = \frac{freq(a)}{N} * \frac{freq(a)}{0.5*(Fleft + Fright)}$$

Здесь

Freq(a) – частотность (количество появления) данного словосочетания (пары) *a*,

N2 - общее число словосочетаний (пар).

Fleft - частота (т.е. количество вхождений в тексте) левой леммы как отдельного слова,

Fright - частота (т.е. количество вхождений в тексте) правой леммы как отдельного слова.

В работе [11] введена метрика, оптимизирующая функцию максимального правдоподобия (*log-likelihood*) в предположении о биномиальном характере функции распределения совместной встречаемости слов.

$$\begin{aligned} loglike = {} & a * log(a+1) + b * log(b+1) \\ & + c * log(c+1) + d * log(d+1) \\ & - (a+b) * log(a+b+1) \\ & - (a+c) * log(a+c+1) \\ & - (b+d) * log(b+d+1) \\ & - (c+d) * log(c+d+1) \\ & + (a+b+c+d) * log(a+b+c+d+1) \end{aligned}$$

Здесь

a - частотность данного словосочетания (пары),

b - суммарная частотность других (отличных от данной) пар с той же самой левой леммой,

c - суммарная частотность других пар с той же самой правой леммой,

d - суммарная частотность пар, отличных от данной и не попадающих в категории (*b*) и (*c*).

В некоторых работах предложены метрики, которые предназначены помочь в выделении словосочетаний большей длины.

Так, K. Frantzi и S. Ananiadou [12] введена метрика C-Value, поощряющая отбор словосочетаний большей длины, которые не входят в состав других словосочетаний.

$$\text{C-Value(a)} = \begin{cases} \log_2 |a| * freq(a) \quad , & \textit{если не вложен} \\ \log_2 |a| - \frac{1}{P(T_a)} * \sum_{b \in T_a} freq(b) \end{cases}$$

здесь *a* – кандидат в термины, $|a|$ - длина словосочетания, измеряемая в количестве слов, *freq(a)* – частотность *a*, T_a – множество словосочетаний, которые содержат *a*, $P(T_a)$ – количество словосочетаний, содержащих *a*.

3. Методы, используемые при формировании состава Общественно-политического тезауруса

Методы, указанные в п.2, ориентированы на выделение устойчивых словосочетаний любого синтаксического вида (например, «глагол+существительное»), так как служат основой для решения широкого круга задач – разрешение многозначности при оптическом распознавании текста, генерации перевода и т.п.

Для задачи формирования тезауруса можно учесть дополнительные обстоятельства.

Анализ имеющихся тезаурусов показывает, что основную массу тезаурусных единиц составляют слова-существительные, а также словосочетания из двух-трех слов. Наиболее часто структура словосочетаний основывается на зависимых от главного существительного прилагательных и существительных в родительном падеже. Оценки, произведенные в работе [3] по терминологии [7], показывают, что термины другой структуры, например, с предлогами или союзами составляют менее одного процента от общего числа терминов, включенных в тезаурус [7].

3.1. Первоначальное наполнение тезауруса

Для разных типов автоматически выделяемых словосочетаний уровень синтаксически неправильных словосочетаний, приходящихся на каждое правильное словосочетание, различен. Для словосочетаний типа «прилагательное + существительное» уровень шума минимален и составляет несколько процентов, что связано с тем, что большинство прилагательных относятся к следующему за ними существительному и согласованы с ним в роде, числе и падеже. Для словосочетаний, внутри которых содержатся предлоги, уровень шума очень велик, так как употребление конструкций с предлогами в русском языке наиболее свободно.

Понятно, что далеко не каждое словосочетание, даже синтаксически правильно выделенное по тексту, может оказаться термином и поэтому поступающий массив словосочетаний должен просматриваться экспертом, который и определяет, является ли данное словосочетание термином и нужно ли его вносить в тезаурус.

Большой процент неоднозначности при выделении предложных конструкций и относительно небольшое количество терминов среди таких конструкций обусловили выбор следующих типов словосочетаний для предъявления их эксперту [3] (обозначим A - прилагательное, N - существительное):

N	существительное
A+N	согласованные прилагательное + существительное
N+N	существительное + существительное в род. падеже
A+A+N	согласованные прилагательное + прилагательное + существительное
N+A+N	существительное + согласованное прилагательное + существительное в родительном падеже

Для описания возможности образования терминов с прилагательными и именными группами в родительном падеже был создан специальный словарь сочетаемости (в определенной степени аналогичная система учета сочетаемости слов используется в программе TextAnalyst [15]).

На основе категорий, приписанных словам, работают правила, которые приписывают словосочетанию категорию "+" или "-" . Категория "+" для словосочетания означает, что словосочетание будет предъявляться эксперту, "-" - словосочетание эксперту предъявляться не будет. Категория словосочетания со словом категории "0" зависит от категорий других слов, входящих в словосочетание.

Обозначим G - группа "прилагательное + существительное", примеры правил:

A(-)+N(-)=G(-)	*важная проблема*
A(+)+N(-)=G(+)	*внешнеполитическая деятельность*
A(-)+N(+)=N(+) (G=N)	*вчерашняя продажа*

Словарь сочетаемости в настоящее время насчитывает около 30000 входов. Считается, что всякое новое относительно словаря слово, появившееся в тексте, имеет категорию "+". По отношению к зависимой конструкции в родительном падеже новое существительное имеет категорию "-".

В 1994-1997 гг. в системе автоматизированной разработки тезауруса было обработано около 50 тысяч официальных документов, что составляет порядка 200 Мбайт текстовой информации. Выявлено более 300 тысяч слов и словосочетаний, которые были просмотрены экспертами. На основе этих словосочетаний была создана первая версия Общественно-политического тезауруса - около 28 тысяч тезаурусных входов (дескрипторов и синонимов, без собственных имен).

3.2. Пополнение существующего тезауруса

После реализации первоначального наполнения, постепенно выяснилось, что метод, описанный в п.3.1, равно как и методы, описанные в п.2, требуют доработки для задачи пополнения уже существующего большого тезауруса.

Действительно, для получения новых «хороших» терминов приходится обрабатывать большие массивы новой информации, при этом увеличивается относительный процент «мусора», что увеличивает трудоемкость работы экспертов.

Поэтому на практике, описанная в п.3.1 технология применяется в УИС РОССИЯ в случаях, когда необходимо подключить в обработку большое количество однотипных текстов смежной предметной области (по избирательной, по социально-экономической тематике).

Та же проблема наблюдалась при формировании подтезаурусов по предметным областям, имеющим значительное пересечение с ранее описанной (например, техническая документация по бортовым системам самолетов), но содержащей большое количество специальных терминов, в том числе в виде предложных групп, «длинных» словосочетаний. К тому же трудно обеспечить однородный корпус текстов по сравнительно узким предметным областям. При этом имеется путь создания своего словаря сочетаемости для каждой предметной области, однако, как мы видели, характерные размеры такого словаря достаточно велики, создание и ведение такого словаря достаточно трудоемко.

Был разработан алгоритм [2], основанный на учете структуры связного текста, ориентированный на обеспечение большей точности выделения терминов типа предложных словосочетаний, а также многословных словосочетаний.

3.3. Алгоритм сборки терминологических словосочетаний на основе внутренних частот употребления словосочетаний в текстах

Если автор текста использует некоторый термин, как отдельную единицу изложения, опирается на этот термин в своем изложении, то именно в этом тексте слова термина рядом будут встречаться чаще, чем в разбивку.

Для того, чтобы это обнаружить при обработке текста для каждого слова (существительного, прилагательного) запоминается его непосредственный слово-сосед и слова-соседи в текстовом окне, заданной величины. Создаются таблица непосредственных соседей и таблица соседей в текстовом окне, вычисляется частотность встречаемости пар слов.

Далее предполагается, что если пара слов встречается как непосредственные соседи более чем в половине случаев их появления в одном и том же текстовом окне, то это свидетельство того, что эта пара слов в совокупности служит автору опорной точкой, то есть представляет собой термин или фрагмент термина.

В этом случае происходит склейка пары слов в единую терминологическую единицу, и таблицы пересчитываются так, как будто бы эта единица была известна с самого начала, до начала обработки текста, что дает возможность и дальше наращивать термин, получая единицы длиной 3 и более.

Приведем технические детали алгоритма.

Основной таблицей является Таблица №1, в которой хранятся сами элементы, их эффективная частотность *Freq1*.

В начале Таблица №1 заполняется леммами и многословными терминами, выбираемыми из словарей терминов тезауруса или иными элементами, собираемыми другими алгоритмами (например, Фамилия+Имя+Отчество). *Freq1* полагается равной реальной частотности элемента в тексте.

Поддерживается Таблица №2 - таблица непосредственных соседей элементов из Таблицы №1.

Допустимыми являются пары вида N+N, A+N (здесь свойство «N» - «существительного» переносится и на словосочетание). Лингвистические ограничения (согласование) в настоящее время не учитываются.

Для каждой пары поддерживается эффективная частотность *Freq2*. В зависимости от назначения может допускаться включение предлогов, иных разделителей, устойчивых оборотов и т.п. Дополнительным ограничением является нахождение в рамках одного предложения.

Аналогичную структуру имеет Таблица №3 - таблица лексических связей, которые устанавливаются между парами элементов на расстоянии, не превышающем заданного предела k, измеряемого в элементах (в настоящее время k=4).

Здесь пары элементов могут находиться в разных предложениях, но не могут находиться в разных абзацах.

Производится цикл по парам элементов Таблицы №2, начиная с пары, имеющей *Argmax(Freq2)*.

Цикл останавливается на значении $Freq2 < 3$, если $ArgMax(Freq1) >= 10$, и на значении $Freq2 < 2$, иначе.

Для каждой пары элементов из Таблицы №2 проверяется следующее условие:

$$Freq2\ (Item_i\ ,\ Item_j) \ >\ 0.5 * Freq3(\ Item_i\ ,\ Item_j),$$

здесь *Freq3* – эффективная частотность лексической связи между элементами $Item_i$ и, $Item_j$.

Если условие не выполняется, то выбирается следующая пара, если выполняется, то производится сборка нового словосочетания.

При сборке производится склейка элементов $Item_i$ и $Item_j$, образуется новый элемент $Item_0$, который записывается в таблицу №1.

При этом, производится пересчет эффективных частотностей элементов в Таблице №1:

$$Freq1(Item_0) := Freq2(Item_i\ ,\ Item_j)$$

$$Freq1(Item_i) := Freq1(Item_i) – Freq1(Item_0)$$

$$Freq1(Item_j) := Freq1(Item_j) – Freq1(Item_0)$$

Также производится пересчет частотностей в таблицах №2 и №3, затем переход к следующему словосочетанию. При практической реализации таблицы №2 и №3 можно объединить в одну.

Примерами полученных таким образом новых терминов для нормативных актов за январь-июнь 2003 г. являются: «*закон об обязательном страховании гражданской ответственности владельцев транспортных средств», «задолженность по обязательным платежам в федеральный бюджет»,* уверенно собираются полные наименования всех органов власти РФ и т.п. В авиационной предметной области – «*положение дежурство на аэродроме», «уничтожение самолета противника», «дежурство в воздухе», «ввод в бой», «выход в боевое соприкосновение»* и др. В предметной области «Выборы» были получены такие термины как *«член избирательной комиссии с правом совещательного голоса», «исполнительный орган местного самоуправления», «выборы главы местного самоуправления».*

3.4. Управление процессом формирования базы терминологических словосочетаний

В процессе решения задачи пополнения уже существующего терминологического ресурса, либо решая задачу извлечения терминологии узкой предметной области по корпусу текстов, относящихся к широкой предметной области, становится актуальным фильтрация набираемых словосочетаний.

Для этой цели мы в своей работе применяем простые схемы, называемые «правилами сочетаемости моделей».

Каждому анализируемому слову или словосочетанию может быть присвоен дополнительный символьный атрибут, так называемая «модель».

Кроме того задаются правила сочетаемости моделей – если имеется строка, описывающая, что элемент модели M1 может следовать за элементом модели M2 (при выполнении тех или иных дополнительных критериев), то и новое словосочетание получает атрибут указываемой в правиле модели M3.

Начальные атрибуты моделей определяются либо при сопоставлении со списками, которым приписываются соответствующие им модели, либо по результатам специальных автоматических процедур (например, для имен, должностей и т.п.).

Производя необходимую сортировку по значениям атрибутов моделей можно быстро обрабатывать получаемые результирующие списки словосочетаний.

Соответственно, задавая или не задавая правила сочетаемости моделей, можно очень просто эффективно управлять составом набираемых словосочетаний, например:

- исключить какие-то классы слов и словосочетаний из набора словосочетаний (актуально для исключения имен собственных);
- наоборот, собирать в одну группу те или иные словосочетания (например, объединяя все словосочетания, содержащие *«Россия»*, *«РФ»*, *«российский»* - для группирования всех полных наименований министерств, ведомств и т.п.);
- учитывать порядок следования элементов разных классов, что часто может уменьшать ошибку сборки словосочетаний.

С помощью моделей можно аппроксимировать правила сочетаемости, описанные в п.3.1.

Пусть

Обозна-чение модели	Описание
00	модель по умолчанию, присваивается изначально словам, неизвестным словарях
01	слова из словаря «+»
02	слова из словаря «+» для родительного падежа
03	слова из словаря «0»
04	слова из словаря «-»
99	имена, отчества, фамилии, топонимы и т.п.

Тогда правила сочетаемости могут быть заданы следующим образом (левый столбец – для первого элемента нового словосочетания, верхняя строка – для второго элемента, на пересечении строки и столбца указан атрибут модели результата).

	00	01	02	03	04	99
00	00	01	02	03	03	--
01	01	01	01	01	03	--
02	00	01	02	--	--	--
03	00	01	02	--	--	--
04	03	03	--	--		--
99	--	--	--	--	--	--

Варьируя правила сочетаемости можно быстро изменять и состав результирующих списков словосочетаний, тем самым оценивая и важность тех или иных правил.

4. Оценка эффективности методов выделения терминологических словосочетаний

Для своих оценок методов отбора терминов мы используем текстовый корпус за шесть месяцев 2003 гг. из состава УИС РОССИЯ – по текстам газет: «Ведомости», «Известия», «Независимая газета», включая приложения «НГ ExLibris», «Независимое военное обозрение», «НГ Религии», «Комсомольская правда», «Аргументы и факты» (ежемесячно порядка 20 Мбайт, 5000 статей) и текстам нормативных актов федерального уровня из коллекции НТЦ «Система» (ежемесячно в среднем 3-4 Мбайта, 300-500 документов).

4.1. Процедура оценки

Мы рассматривали следующие методы:

Обозначение метода	Пояснения
PAIRS	метод суммирования частотностей пар рядом стоящих слов без учета лингвистических ограничений
PAIRS.LL	результат PAIRS упорядочивается в соответствии с критерием «log-likelihood»
PAIRS.MI	результат PAIRS упорядочивается в соответствии с критерием «mutual information»
OLDTERM--	метод, описанный в п.3.1 (пары и тройки без предлогов), без

Обозначение метода	Пояснения
	использования словарей сочетаемости
OLDTERM.++	метод, описанный в п.3.1 с использованием словарей сочетаемости
C-VALUE	метод «C-value» - сборка словосочетаний любой длины в соответствии с механистическим критерием
TERMS--	метод, описанный в п.3.3, без использования словарей
TERMS++	метод, описанный в п.3.3, с использованием простых словарей отсечения – имена и географические названия
TERMS+THES	метод TERMS++, причем в качестве начальных словарей используются термины из Общественно-политического тезауруса РуТез

Для всех методов производилась единообразная процедура.

Обрабатывался каждый документ, получившиеся списки словосочетаний для каждого метода были сведены в один, при этом суммировалось количество документов и частотность (эффективная частотность). Списки полученных словосочетаний упорядочивались в соответствии с убывающей суммарной частотностью для PAIRS и OLDTERM, по убыванию суммы оценки для методов PAIRS.LL, PAIRS.MI, C-VALUE, по убыванию суммарной эффективной частоты для всех видов TERMS. При этом не учитывались словосочетания, которые входили на рассматриваемой коллекции только в один документ.

Полученные результирующие списки сопоставлялись со списком словосочетаний, образованным объединением терминов Общественно-политического тезауруса РуТез (для чистоты эксперимента бралась версия от 27.12.2002), встречавшихся в данной коллекции (4187 терминов, получены из результатов TERMS+THES) и списка, образованного согласно п.4.2.

4.2. База оценки

Оценка эффективности разных методик формирования базы терминологических словосочетаний затруднена, так как требует принятия решения экспертом о том, является ли данное словосочетание термином или нет.

При больших объемах баз отобранных словосочетаний проведение такого рода оценок весьма трудоемко и не освобождает от субъективности экспертов [12], которые должны иметь одинаковое представление о границах рассматриваемой предметной области. Менее трудоемко сравнивать, накладывая получившиеся словосочетания на уже существующие тезаурусы, здесь необходимо учитывать неполноту любого тезауруса.

Мы постарались провести детальный анализ полученных результатов на коллекции нормативно-правовых документов РФ (НТЦ «Система») с датой выхода с 1 января по 31 июня 2003 г., полученных по 9 июля 2003 г. Всего 2415 документов суммарным размером 16,5 Мбайт.

Для этой цели мы произвели объединение верхних частей списков словосочетаний (первых 2000-5000 строк) полученных разными методами, отфильтровали термины, входящие в Общественно-политический тезаурус. Все остальные были просмотрены экспертами для определения терминологических словосочетаний, которые в настоящее время не входят в тезаурус.

Следует еще раз отметить, что реально использующийся большой терминологический ресурс может развиваться только в случае соблюдения некоторого количества общих принципов. Под большим терминологическим ресурсом понимается такой, в котором уже ни один из экспертов, ведущих тезаурус, не может по памяти, без поверки по базе данных, уверенно сказать - входит предъявляемый термин в тезаурус или нет.

Общественно-политический тезаурус активно используется для решения прикладных задач в широкой предметной области (текстов правовых документов и текстов СМИ общественно-политической направленности): тематического анализа текста, автоматической рубрикации по сопровождаемым рубрикаторам, расширения запроса при информационном поиске, анализа результатов информационного поиска.

К ограничивающим принципам используемым при формировании тезауруса относятся [5]:

- первоочередное включение в тезаурус терминов, которые необходимы для повышения качества результатов решаемых прикладных задач, прежде всего увеличения среднего покрытия текстов терминологическими элементами тезауруса, разрешения многозначности при автоматической обработке;
- системность включения терминов, что требует включения группы взаимосвязанных терминов при «вторжении» новую подобласть.

Так как предметная область весьма широка, то обеспечивая хорошее среднее покрытие описанной терминологией практически каждого текста правовых документов и материалов СМИ, тем не менее имеется достаточной величины «хвост» терминологических слов и словосочетаний, не включенных в тезаурус, это прежде всего термины специальных областей (науки, медицины, техники), которые сравнительно редко обсуждаются в рассматриваемых текстовых коллекциях.

Кроме того, следует учитывать ограниченность трудовых и организационных ресурсов, задействованных в работе по формированию тезауруса.

Реальная ситуация такова, что в любой момент существует около 3000 терминологических словосочетаний, которые ждут своей очереди быть включенными в тезаурус (быть приписанными к конкретному понятию в понятийной сети). При этом данная «очередь» пополняется из различных по приоритетности обработки источников, некоторые кандидаты находятся в этой очереди годами.

В обсуждаемой задаче эксперты в своей работе по разметке словосочетаний-кандидатов использовали следующие пометки, отражающие определенную классификацию получившихся словосочетаний:

Пометка	Примечание
t	Эксперт уверен, что словосочетание может быть включено в тезаурус (*«военное назначение»* - дополнительно снимает многозначность с обоих слов по отдельности). Это не значит, что помеченные так словосочетания реально будут включены в тезаурус, так как скорее всего им придется ждать своей очереди
!	Скорее термин чем не термин. Дело в том, что принять точное решение можно, только сопоставив включаемый термин с уже существующими, оценив влияние рассматриваемого изменения на ВСЕ возможные сферы приложения Общественно-политического тезауруса. Эксперт не уверен, что предъявленное словосочетание будет улучшать использование тезауруса для всех приложений.
?	Эксперт еще более не уверен. Слишком большая близость к свободным словосочетаниям в контексте широкой предметной области (*«длительное пользование»*, *«дневная форма»*, *«заседание совета директоров»*). Возможно, имеются некоторые приложения или узкие предметные подобласти, где наличие предъявленного словосочетания в составе тезауруса может помочь в автоматической обработке. Но издержки слишком велики.
a	Эта пометка использовалась для иллюстративных целей – помечались словосочетания обозначающие фрагменты документов, действия над документами. Такие словосочетания, практически пустые в смысле классической терминографии, весьма полезны, например, для классификации документов по рубрикам типа «Изменение нормативного акта …». Могут быть включены в специальные подобласти тезауруса для автоматической обработки.
e	Ошибка. Словосочетание никогда не будет включено в тезаурус. Может быть подразделено на два случая: «пустых» словосочетаний (*«альтернативное условие»*), «осколков» более длинных реальных терминов (*«безопасность ООН»* от *«Совета Безопасности ООН»*) и явных синтаксических ошибок сборки (например, куски таблиц – *«ВЭД | наименование»*).
g	Топонимы.
h	Ошибки, включающие в себя топонимы.
i	Имена, отчества персон, а также словосочетания куда входят имена, но не являющиеся синтаксической ошибкой (*«Андрей Алексеевич»*), например, плюс фамилия.
j	Ошибки, включающие в себя имена
o	Наименование конкретных организации *«НПО Энергомаш»*, а также организаций, определяемых топонимами, например, *«барнаульский завод»*.

Подчеркнем, что наличие гаммы оценок пригодности отнесения того или иного словосочетания в тезаурус определенным образом свидетельствует об отсутствии абсолютных правил автоматического выделения терминологических словосочетаний.

Для проведения оценки, в качестве «хороших» терминологических словосочетаний далее использовались словосочетания, получившие пометки (t) и (!).

Одновременно, анализ ошибок показывает, что количество неправильных словосочетаний можно достаточно просто уменьшить, запретив собирать (используя аппарат моделей, описанный в п.3.4) словосочетания, содержащие:

- имена, отчества, известные фамилии;
- топонимы - собственные имена, а также производные от топонимов («район», «область», «город» и т.п.). Здесь требуется соблюдать осторожность, так как существует много терминов с указанными словами. Лучше всего сначала собрать их в отдельный список с помощью специальных моделей, а затем дополнительно отфильтровать;
- пустые слова, например, названия месяцев, валют, числительных (*«млн», «млрд», «тысяча»*), собственно, пустые в предметной области слова (*«указанный», «необходимый», «необходимость»,*

«соответствующий» и т.п.). В целом, таких оказалось сравнительно немного – два-три десятка.

4.3. Оценка качества сборки словосочетаний

Анализируемые методы (см. п.4.1) разделяются на две группы.

Методы группы PAIRS и OLDTERM предназначены для первоначального «быстрого» заполнения терминологического ресурса объектами достаточно простой природы. Методы группы C-VALUE и TERMS ориентированы на выделение более сложных объектов, более длинных словосочетаний. Платой за решение этой более сложной задачи должен являться и больший шум в результатах обработки.

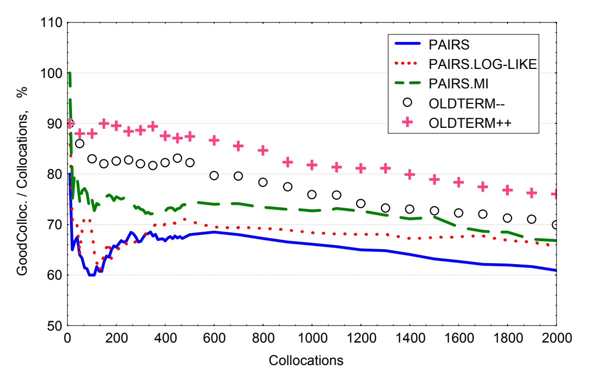

Рис.1. Оценка эффективности нахождения коротких терминологических словосочетаний

На Рисунке 1 представлено сравнение распределений «хороших» терминоподобных словосочетаний среди всех найденных для методов группы PAIRS и OLDTERM.

В этом и следующих рисунках по горизонтальной оси отложено количество словосочетаний от начала результирующего списка, отсортированных в порядке убывания суммарной частотности (эффективной частотности). По вертикальной оси отложено процентное отношение количества хороших словосочетаний среди всех с начала списка.

Как можно видеть на Рисунке 1, для рассматриваемой коллекции документов при сборке коротких словосочетаний преимущество имеют методы OLDTERM, использующие локальную лингвистическую информацию. Модификации метода PAIRS по пересортировке результатов в соответствии с механистическими критериями приводят к улучшению по сравнению с базовым методом, но эти улучшения не носят принципиального характера.

Лучшие результаты достигаются при использовании словарей отсечения (OLDTERM++), прежде всего отсечения имен людей и топонимов.

На Рисунке 2 представлено сравнение распределений «хороших» терминоподобных словосочетаний среди всех найденных для методов группы методов групп C-VALUE и TERMS. Для облегчения сопоставления с результатами методов PAIRS и OLDTERM на рисунке 2 повторены данные для метода OLDTERM++.

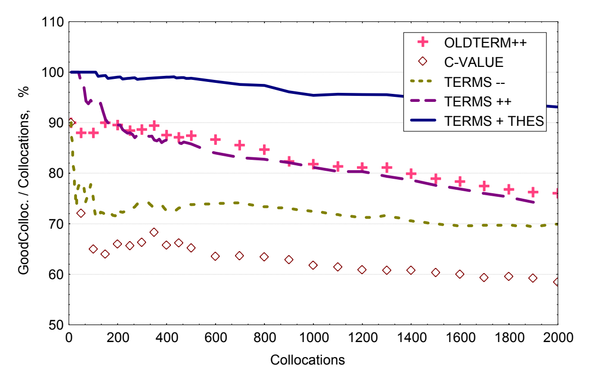

Рис.2. Оценка эффективности нахождения в классе «любых» терминологических словосочетаний

Наблюдается преимущество методов группы TERMS по сравнению с C-VALUE. Здесь необходимо отметить, что авторами метода C-Value K. Frantzi и S. Ananiadou было предложено большое количество модификаций указанного алгоритма для повышения качества его результатов. Рассмотрение всех этих модификаций выходит за рамки данной работы.

Результаты метода TERMS++ - с использованием словарей отсечений - значительно лучше чем у простого метода TERMS.

4.4. Сборка «длинных» словосочетаний

Результаты по методу TERMS+THES приведены на рисунке 2 для иллюстрации, так как на первом же шаге алгоритма определяются словосочетания, описанные в тезаурусе.

Достоинства метода TERMS+THES проявляются в задаче нахождения терминологических словосочетаний длиной более чем три слова - не считая разделителей и предлогов (см. Рисунок 3).

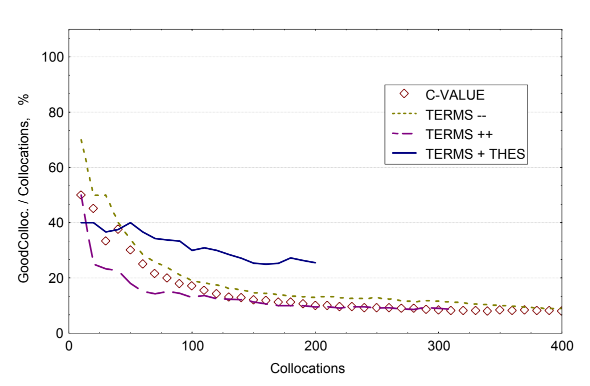

Рис.3. Оценка эффективности нахождения «длинных» терминологических словосочетаний

Сборка таких словосочетаний весьма сложна, даже лучший из применяемых нами методов (TERMS+THES) дает точность около 30%, но и это значительно больше, чем при использовании других подходов. Преимущество TERMS+THES может быть объяснено за счет отсутствия брака при сборке коротких словосочетаний, которые все методы стараются «удлинить» - другие методы «удлиняют» наряду с хорошими также плохие словосочетания.

4.5. Нахождение терминов тезауруса

Еще одной интересной характеристикой методов сборки терминологических словосочетаний является полнота нахождения терминов Общественно-политического тезауруса разными методами:

полнота нахождения терминов тезауруса	2000 слвсч.	Всего слвсч.	Кол-во слвсч.
PAIRS	16,8%	68,9%	25449
PAIRS.LL	18,1%	68,9%	25449
PAIRS.MI	18,4%	68,9%	25449
OLDTERM--	19,8%	75,3%	18759
OLDTERM++	21,7%	63,2%	11271
C-VALUE	14,7%	68,6%	34965
TERMS--	17,3%	21,4%	2845
TERMS++	20,5%	24,2%	2368
TERMS+THES	*32,5%*	*100%*	*6567*

Приведенная таблица содержит процентное отношение найденных терминов Общественно-политического тезауруса каждым из исследуемых методов в первых 2000 выдаваемых словосочетаний, всего найденных словосочетаний, а также общее количество выдаваемых словосочетаний. Как нетрудно видеть, методы группы OLDTERM демонстрируют лучшие показатели.

Методы группы TERMS ориентированы на увеличение точности, поэтому показывая неплохие результаты в начале списка, но не обнаруживают большинства терминов тезауруса (исключая, естественно, реализацию TERMS+THES). Для более полной проверки необходим анализ на больших объемах данных.

4.6. Специальные виды словосочетаний

Большой интерес представляет распределения словосочетаний специального вида, прежде всего наименее «сорного» типа «прилагательное + существительное».

На Рисунке 4 и Рисунке 5 приведены результаты для данного типа словосочетаний (из общего результирующего списка производилась выборка, для которой затем повторялись сделанные ранее оценки.)

На Рисунке 4 результаты приведены для методов групп PAIRS и OLDTERM, на Рисунке 5 – для методов групп TERMS и C-VALUE.

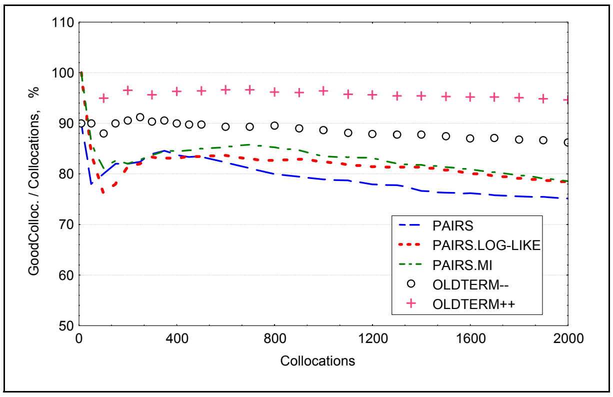

Рис.4. Оценка эффективности сборки пар вида «прилагательное+существительное» для методов групп PAIRS и OLDTERM

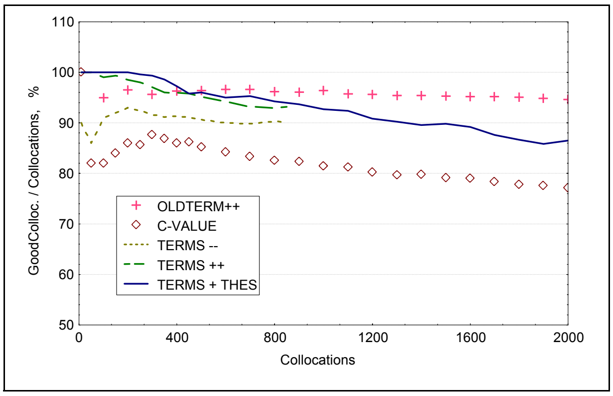

Рис.5. Оценка эффективности сборки пар вида «прилагательное+существительное» для методов групп C-VALUE и TERMS

Как легко видеть, качество выделения терминологических словосочетаний такого типа очень высокое – 80-90%. Следует иметь в виду, что термины такого вида составляют не более двух третей от общего количества словосочетаний-терминов.

Лучшие показатели OLTERM++ перед TERMS+THES в данном случае могут объясняться тем, что в OLDTERM отслеживается согласование между прилагательным и существительным, в TERMS такое требование не накладывалось.

Как и в общем случае (п.4.3), преимущество имеют методы, использующие предварительно накопленные знания.

5. Заключение

Проблема формирования большого терминологического ресурса разделяется на задачу первоначального накопления информации и задачу дальнейшего пополнения ресурса. Для решения этих задач целесообразно использовать разные методы.

Методы, использующие механистические критерии, могут быть использованы для формирования терминологических ресурсов. Но, в целом, они уступают методам, учитывающим лингвистическую информацию.

Существенную роль играет возможность использовать дополнительные лингвистические фильтры, прежде всего словари отсечений для собственных имен, слов-зародышей для «пустых» конструкций.

При сборке словосочетаний большой длины важно использовать списки известных словосочетаний для уменьшения шума из-за неправильно собираемых коротких словосочетаний.

В целом, более предпочтительными оказываются методы, которые могут итерационно использовать накапливаемую на предыдущих этапах информацию – как для уменьшения ошибки сборки коротких словосочетаний, так и для формирования необходимых списков отсечений.

Литература

[1] Антонов А.В., Информационно-поисковая система Galaktika-ZOOM с элементами анализа на гипермассивах информации // НТИ. Сер. 1. – 2001. – № 8. – С.12–21.

[2] Добров Б.В., Лукашевич Н.В., Невзорова О.А., Автоматизированное построение прикладной онтологии: технологические аспекты // Международная IEEE конференция Искусственные интеллектуальные системы (IEEE AIS'02) Геленджик-Дивноморское,– Обработка текста и когнитивные технологии: Сборник (Вып. 7) / Под ред. В.Д.Соловьева – Казань: Отечество, 2002. – С.103-109.

[3] Лукашевич Н.В., Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. - 1995. - №3. - С.21-24.

[4] Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А.С.Нариньяни – М.: Наука – 2002. – Т.2 - С.338-346.

[5] Лукашевич Н.В., Салий А.Д., Тезаурус для автоматического рубрицирования и индексирования: разработка, структура, ведение // НТИ. Сер.2. - 1996. - №1. - С.1-6.

[6] Суперанская А.В., Подольская Н.В., Васильева Н.В., Общая терминология. Вопросы теории. – М.: УРСС – 2003.

[7] Тезаурус информационно-поисковый по законодательству - Мин-во юстиции СССР, Научный центр правовой информации. - М., 1980.

[8] Шелов С.Д., Определение терминов и понятийная структура терминологии. – СПб.: Изд-во СпбГУ – 1998.

[9] Юдина Т.Н., Журавлев С.В., Российский межуниверситетский ресурсный и аналитический центр по гуманитарным исследованиям - Вестник РФФИ. – 1999. - №3 (специальный выпуск), "Наука и информационное общество".

[10] Church K., Hanks P., Word association norms, mutual information, and lexicography // Proceedings of ACL-89. - 1989. - Vancouver, Canada - p.76-83.

[11] Dunning T., Accurate Methods for the Statistics of Surprise and Coincidence // Computational Linguistics – 1993. - 19(1) – p.61-74.

[12] Frantzi K.T., Ananiadou S., Automatic Term Recognition using Contextual Cues // Proceedings of Mulsaic 97, IJCAI, Japan. 8 Contingency tables utilise only the frequency of the word inside the processed text."- 1997.

[13] McKeown K.R., Radev D.R., Collocations // R.Dale, H.Moisl, H.Somers (eds.), A Handbook of Natural Language Processing. - Marcel Dekker, 1998 – Chapter 15.

[14] Smadja F., Retrieving collocations from text: Xtract // Computational Linguistics - 1993. - 19(1) – p.143-177.

[15] TextAnalyst (http://www.analyst.ru).

[16] VIVISIMO (http://www.vivisimo.com).

Automatic Detection of Text Entries for Information Retrieval Thesaurus

B.V.Dobrov[1), 3)], N.V.Loukachevitch[1), 3)],
S.V.Syromyatnukov[2), 3)]

The paper is devoted to constructing and supporting large information retrieval thesauri. The different techniques of collocations extraction are compared. The new techniques are described.

♣ Данная работа частично выполняется при финансовой поддержке грантов РФФИ №01-07-90430 и № 03-01-00472.