Тестирование соответствия тезауруса посредника лексике отдельных коллекций

Казаков Евгений Николаевич
Всероссийский научно-технический информационный центр (ВНТИЦ), Москва
Сомин Николай Владимирович
Институт проблем информатики РАН (ИПИ РАН), Москва
kuznetsy@ccas.ru

Аннотация

Рассматривается проблема сопоставления лексики тезауруса и коллекции электронной библиотеки (ЭБ)в ситуации, когда тезаурус содержится в составе посредника, обеспечивающего взаимодействие пользователя с ЭБ. Предлагаются некоторые количественные характеристики, оценивающие степень согласования лексики коллекции и тезауруса на уровнях словоформ и словосочетаний. Разработаны программные средства доступа к тезаурусу, а также морфологического и синтаксического анализа текстов русского языка, обеспечивающие выделение из текстов слов и именных словосочетаний с последующей их нормализацией. Проведен ряд вычислительных экспериментов по подсчету количественных характеристик для коллекций с помощью двух тезаурусов разного объема. По результатам экспериментов делается вывод о необходимости постановки задачи дополнения тезауруса посредника лексикой вновь подключаемой к нему коллекции.

Результаты работы могут быть использованы при создании посредника и разработке процедур пополнения тезауруса новой лексикой.

1 ПОСТАНОВКА ЗАДАЧИ

Одно из перспективных направлений в создании электронных библиотек (ЭБ) предполагает построение посредника, обеспечивающего единообразное взаимодействие пользователей с многочисленными коллекциями любого типа. Введение посредника между пользователем и ЭБ призвано решить проблему масштабирования взаимодействия, т.е. обеспечить фиксированную (и приемлемую для пользователя) сложность сценария доступа к ЭБ, не возрастающую с ростом числа и разнообразия ЭБ.

Помимо других средств, в состав посредника целесообразно ввести тезаурус, который, являясь единой терминологической базой посредника, обеспечит интеллектуальную помощь в процессе формирования запроса пользователем безотносительно к лексической специфике той или иной коллекции. Однако, для этого тезаурус посредника должен стать интегрированным, т.е. содержать иерархические и синонимичные связи с существенной частью лексики всех связанных с ним коллекций.

Для достижения этой цели могут быть использованы разные стратегии. Каждая ЭБ и коллекция, с которой может взаимодействовать посредник, должна подключаться к нему с помощью процедуры регистрации. При этом возможно, что в составе ЭБ имеется собственный локальный тезаурус или словарь. Поэтому для создания интегрального тезауруса в принципе можно было бы пойти по пути объединения тезаурусов ил словарей коллекций. Однако представляется, что более перспективным будет второй подход, базирующийся на использовании некоего нормативного политематического тезауруса, который дополняется процедурой возможного пополнения лексикой или терминологией локальных тезаурусов коллекций.

Вкратце, второй подход выглядит следующим образом. Если в составе коллекции есть собственный тезаурус, то в процессе регистрации должно происходить его сравнение с тезаурусом посредника с выделением совпадающей и несовпадающей лексики. Если в коллекции нет тезауруса, то с тезаурусом посредника сравнивается лексика коллекции. Процедура сравнения должна оценить степень соответствия тезауруса посредника лексике конкретной коллекции и, соответственно, обосновать решение: либо оставить тезаурус посредника без изменений, либо дополнить его частью лексики регистрируемой коллекции. Процедура формирования и ведения интегрированного тезауруса в общем виде описана в [1] в предположении, что нормативный тезаурус посредника и тезаурус или терминологический словарь коллекции есть в готовом виде. Методика, используемая в данной работе, направлена прежде всего на реализацию такой процедуры в случае, когда коллекция не имеет своего тезауруса или словаря.

Настоящая работа носит предварительно-оценочный характер и призвана на основе ряда вычислительных экспериментов ответить на основные вопросы, связанные с оценкой реалистичности реализации второго подхода, а именно: 1) может ли какой либо из существующих те-

©Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Протвино

заурусов использоваться в качестве нормативного и 2) насколько необходима реализация процедуры пополнения тезауруса, или, может быть нормативный тезаурус сам может служить надежной лексической базой для формулирования запроса к любой текстовой коллекции и, следовательно, процедура его пополнения является излишней.

2 ВЫБОР МЕРЫ СООТВЕТСТВИЯ

Для сравнения лексики тезаурусов посредника и коллекции в [1] предлагались:

- коэффициент совпадения словоформ;
- коэффициент покрытия тезауруса посредника словоформами коллекции;
- коэффициент покрытия словоформ коллекции словоформами тезауруса посредника.

Аналогичные коэффициенты могут быть определены для словосочетаний. С позиции конкретной коллекции наиболее важными являются коэффициенты покрытия словоформ (Kc) и словосочетаний (Kcc) коллекции лексикой тезауруса посредника:

$$Kc = |Skt|/|Sk|,$$

где: $|Skt|$ - мощность множества словоформ, общих для коллекции и тезауруса посредника; $|Sk|$ - мощность множества словоформ коллекции;

$$Kcc = |Ckt|/|Ck|,$$

где: $|Ckt|$ - мощность множества словосочетаний, общих для коллекции и тезауруса посредника; $|Ck|$ - мощность множества словосочетаний коллекции.

Если тезаурус или словарь коллекции отсутствует, то вычисление Sk и Ck может быть выполнено следующим образом. Множество Sk образуют нормированные по роду, числу и падежу существительные и прилагательные в текстах коллекций, а множество Ck - нормализованные по числу и падежу именные словосочетания, образуемые из опорного существительного, дополненного слева прилагательными: а справа - генитивными цепочками.

Однако, при сравнении тезауруса посредника с лексикой текстов коллекций, эти критерии имеют два существенных недостатка. Во-первых, коэффициенты Kc и Kcc не учитывают частоту встречаемости слов или словосочетаний, что не позволяет статистически точно оценить степень соответствия лексики коллекции и тезауруса. О том, что учет частот встречаемостей совершенно необходим говорит и закон Ципфа, согласно которому 90% текста образует половина его лексического состава. Во-вторых, извлекаемые из коллекции словосочетания, в силу чисто синтаксического метода их выделения, могут не точно соответствовать терминам тезауруса, а содержать дополнительные прилагательные или генитивные цепочки, так что полное совпадение термина тезауруса словосочетанию из текстов коллекции оказывается слишком жестким требованием.

Для нивелирования этих негативных факторов были предложены следующие коэффициенты:

Kw - отношение числа вхождений в тексты коллекции значащих слов тезауруса(существительных и прилагательных); к числу вхождений всех значащих слов в текстах коллекции;

Ko - отношение числа вхождений выделенных из коллекции словосочетаний, в которые входит хотя бы один из терминов тезауруса к числу всех вхождений словосочетаний коллекции;

Ko_i $(i = 0, 1, 2, ...)$ - отношение числа вхождений словосочетаний, включающих термин тезауруса к числу всех вхождений словосочетаний коллекции. При этом точность соответствия оценивается разностью i между длинами словосочетаний коллекции и тезауруса: при $i = 0$ имеет место точное равенство словосочетаний; при $i = 1$ словосочетание коллекции длиннее на одно слово, чем словосочетание тезауруса. Отметим, что при таком определении коэффициент Ko является "пределом"последовательности Ko_i при неограниченном увеличении индекса i;

Kv и Kv_i $(i = 0, 1, 2, ...)$ - аналогичны Ko и Ko_i, но дополнительно требуется совпадение опорных существительных сравниваемых словосочетаний.

Эти коэффициенты и были использованы при проведении компьютерных экспериментов. Априори, на основании накопленного авторами опыта обработки текстов, предполагалось, что значения всех коэффициентов (или большинства из них) будут находиться в диапазоне 0.8 - 1, то это свидетельствует о хорошем соответствии тезауруса посредника лексике коллекции и пополнения тезауруса не требуется. С другой стороны, значения коэффициентов менее 0.4 - 0.5 на значительном множестве коллекций говорят о том, что тезаурус либо бедно насыщен лексикой, либо слишком специализирован и использование его в качестве нормативного является проблематичным.

3 НОРМАТИВНЫЙ ТЕЗАУРУС ПОСРЕДНИКА

В качестве нормативного тезауруса посредника был принят политематический тезаурус Всероссийского научно-технического информационного центра (ВНТИЦ), который был сформирован динамическим компьютеризованным методом [2] с участием коллектива специалистов по различным областям знаний на основе обработки текстов рефератов к отчетам о научно-исследовательских и опытно-конструкторских работах, а также к кандидатским и докторским диссертациям, относящихся ко всем областям науки и техники. Лексика тезауруса ВНТИЦ покрывает тематику всех рубрик Государственного рубрикатора научно-технической информации (ГРНТИ) [3], что характеризует тезаурус ВНТИЦ как тезаурус универсальный по тематике.

В тезаурусе ВНТИЦ фиксируются следующие семантические связи: выше по иерархии, ниже по иерархии и синонимия. Полный тезаурус содержит около 225 тыс. терминов (лексических единиц) и свыше 145 тыс. семантических связей. В машинных экспериментах использовалась несколько суженная лексически версия тезауруса, содержащая 151700 терминов, состоящих из 58009 слов.

Для сравнения использовался также тезаурус

культурно-социальных терминов, разработанный в ИНИОН и содержащий 9724 термина, состоящих из 6234 слов. Следует отметить, что указанные тезаурусы имеют довольно скромную общую часть: 2332 термина и 2484 слова.

4 МЕТОДИКА ЭКСПЕРИМЕНТОВ

Для работы с тезаурусами в среде Windows-95 разработаны специализированная БД, ориентированная на работу с текстовыми словарями, а также программа загрузки тезауруса в БД из текстовых файлов. Структура БД включает словарь терминов и слов тезауруса и две реляционные таблицы: таблицу синтаксических связей между словами терминов и таблицу семантических связей между терминами.

Обработка текстов коллекций осуществлялась следующим образом. Для каждого слова коллекции выполнялся морфологический анализ, для чего использовалась разработанная в ИПИ РАН система морфологического анализа русского языка [4]. Определялась часть речи; для прилагательных, причастий и существительных определялись род, число, падеж и некоторые другие специфические параметры.

Затем, путем "склеивания"рядом стоящих словоформ по алгоритмам полного согласования и присоединения генитивных цепочек, определялось "естественное"словосочетание. Таким образом, словосочетание состоит из существительного (которое мы будем называть опорным), перед которым может стоять несколько прилагательных или причастий, а после него - генитивная цепочка из существительных в родительном падеже, возможно также осложненных прилагательными или причастиями. Например, в словосочетании "основной показатель внешнеактивной деятельности предприятия"существительное "показатель"является опорным; оно осложнено слева прилагательным "основной", а справа двумя генитивными цепочками "внешнеактивная деятельность"и "предприятие". Далее, ради упрощения алгоритма сопоставления, все словоформы словосочетания преобразуются в канонический вид, так что вышеприведенное словосочетание приобретает вид: "основной показатель внешнеактивный деятельность предприятие". Приведение к такому каноническому виду выполнялось и для всех терминов тезаурусов.

Наконец, для подсчета коэффициентов Ko, Ko_i, Kv и Kv_i, выполнялось сопоставление каждого из найденных в тексте словосочетаний с терминологией тезаурусов. Для подсчета коэффициента Kw выполнялось сопоставление словарного состава словосочетаний со словарным составом тезаурусов.

Программы подсчета коэффициентов и программное обеспечение БД реализованы на языке C в среде Microsoft Visual C.

5 РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

С помощью тезауруса был осуществлен ряд экспериментов по обработке текстовых баз данных.

Исследовалось два контрольных примера, представляющих собой коллекции текстов на русском языке: 1) пример "S", содержащий 33520 терминов и 50550 значимых слов; и 2) пример "D", содержащий 65112 терминов и 140220 значащих слов. Если первая коллекция содержит общественно-политические тексты, то тексты коллекции 2) являются аннотациями докторских диссертационных работ практически по всем специальностям ВАК. Полученные результаты (в %) представлены в следующей таблице (заметим, что значения коэффициента Kv_0 не приводятся, поскольку $Kv_0 = Ko_0$).

	ИНИОН		ВНТИЦ	
	"S"	"D"	"S"	"D"
Kw	42	41	53	57
Ko_0	21	16	34	30
Ko_1	30	26	45	45
Ko_2	33	34	49	55
Ko_3	34	39	50	61
Ko	35	43	51	65
Kv_1	24	22	38	35
Kv_2	26	27	41	42
Kv_3	27	30	41	45
Kv	27	33	41	48

6 ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Проведенные эксперименты показали, что как объем и качество тезауруса, так и тематическая направленность коллекции существенно влияют на покрытие лексики ЭБ. Легко видеть, что если тезаурус ИНИОН на обоих тестах дает примерно одинаковые результаты, то тезаурус ВНТИЦ лучшие результаты дает на текстах научно-технической направленности.

Кроме того, всякое ужесточение критериев сопоставления терминов (требование совпадения опорных существительных, уменьшение числа i несовпавших словоформ) тезауруса и коллекции существенно снижает процент найденных в тезаурусе терминов.

Приведенные результаты показывают, что тезаурус ИНИОН вряд ли может рассматриваться, как основа для построения лексической базы посредника. Другое дело тезаурус ВНТИЦ: его политематичность и существенно лучшие результаты, в большинстве случаев превышающие или близкие к порогу в 50%, полученные причем как на общественно-политических, так и научно-технических текстах дают основание полагать, что это тезаурус может служить терминологической базой, на основе которой возможно построение интегрального тезауруса посредника.

Однако, как видно из приведенной таблицы, ни один из коэффициентов не превосходит априорный порог "полноты тезауруса"0.8. Это говорит о том, что даже при использовании тезауруса ВНТИЦ в коллекции остается значительное терминологическое наполнение, не охватываемое тезаурусом.

Поэтому следует признать актуальной решение проблемы "остаточной лексики", т.е. совокупности лексико-терминологических единиц, не входящей в тезаурус посредника, но существенной для индексации документов коллекции. Необходимо найти процедуры выделения такой "остаточной"лексики, а также создать программные

средства для пополнения этими лексическими единицами тезауруса посредника и предоставления их пользователю посредника при формулировании запросов.

Список литературы

[1] Казаков Е.Н. Формирование и ведение тезауруса в составе посредника между пользователями и сетью электронных библиотек// Труды Первой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Санкт-Петербург, 1999. с. 85-88.

[2] Казаков Е.Н., Копылов В.А. Динамический способ построения информационно-поисковых тезаурусов//Научно-техническая информация, сер. 2, Москва, ВИНИТИ, 1974, N 5, с. 20-28.

[3] Государственный рубрикатор научно-технической информации (ГРНТИ), Москва, 4-е изд., 1992, 248 с.

[4] Сомин Н.В., Соловьева Н.С., Соловьев С.В. Система рубрикации текстовых сообщений. //Труды Международного семинара Диалог'98 по компьютерной лингвистике и ее приложениям: В 2 т. Т. 2./Под ред. А.С. Нариньяни. - Казань: ООО "Хэтер", 1998. - С. 574-581.