ДВУЯЗЫЧНЫЙ ИНФОРМАЦИОННЫЙ ПОИСК НА ОСНОВЕ АВТОМАТИЧЕСКОГО КОНЦЕПТУАЛЬНОГО ИНДЕКСИРОВАНИЯ

Н.В.Лукашевич, Б.В.Добров
Научно-исследовательский вычислительный центр МГУ им.М.В.Ломоносова;
АНО Центр информационных исследований
339, НИВЦ МГУ, Воробъевы горы, Москва, 119899
{dobroff, louk}@mail.cir.ru

1. Введение

Важной задачей современного информационного поиска является разработка многоязычных информационных систем, в которых запрос составляется на одном языке, а найденные документы могут быть как на языке запроса, так и на других языках информационной системы. В последнее время проблема многоязычного поиска стала одной из приоритетных в информационном поиске, например, в рамках конференции TREC выделено соответствующее направление [4]. С этой работой тесно взаимодействуют исследователи, объединенные европейской программой CLEF (Cross-Language Evaluation Forum) [1, 5], а также японской программой NTCIR [6].

Задачи поиска документов в многоязычных текстовых коллекциях решаются уже в течение нескольких десятилетий. Для решения таких задач создавались многоязычные информационно-поисковые тезаурусы [14], в которых для каждого дескриптора тезауруса были сформулированы его варианты на нескольких языках. Однако традиционные многоязычные тезаурусы создавались как вспомогательные средства для ручного индексирования/поиска специалистами-индексаторами.

Возникновение громадных электронных коллекций потребовало разработки средств автоматической обработки и поиска документов в многоязычных коллекциях. В настоящее время для решения задачи поиска в многоязычной среде активно исследуются пословные методы, основанные на использовании параллельных и сравнимых корпусов, традиционных двуязычных словарей [5].

Пословным методам информационного поиска обычно противопоставляются методы так называемого концептуального индексирования, при которых для каждого текста автоматически строится не пословный, а понятийный индекс, в котором синонимичные выражения представлены одним и тем же понятиям, многозначность текстовых выражений отражена различными понятиями, и имеется возможность использования отношений между понятиями для автоматического расширения запроса [7, 9, 10]. Реализация таких подходов затруднена необходимостью разработки специальных понятийно-ориентированных лингвистических ресурсов – тезаурусов (онтологий).

Именно разработка такого многоязычного ресурса для автоматического концептуального индексирования документов на европейских языках ставилась в проекте EuroWordNet [2]. В рамках этого проекта для ряда языков были созданы лингвистические ресурсы общего назначения, которые нужно еще дополнять при работе в конкретных предметных областях, что является отдельной проблемой.

В нашем докладе мы опишем основные принципы разработки русско-английского Тезауруса по общественно-политической жизни [15], который в настоящее время в русской части имеет 67 тысячи терминов, в английской части 60 тысяч терминов, представленных как иерархическая сеть 28 тысяч понятий. На основе Тезауруса производится автоматическое концептуальное индексирование русских и английских документов, при этом производится автоматическое разрешение многозначности терминов. Построенный концептуальный индекс

позволяет выполнять поиск одновременно русских и английских документов по запросу на русском или английском языке в Университетской информационной системе РОССИЯ (УИС РОССИЯ, http://www.cir.ru) [13].

2. Традиционные тезаурусы и автоматическое концептуальное индексирование: тезаурус EUROVOC

Проблемы применения традиционных информационно-поисковых тезаурусов в процессе автоматической обработки текстов рассмотрим на примере тезауруса EUROVOC – тезаурусе органов Европейского Союза. Русскоязычная версия тезауруса EUROVOC подготовлена сотрудниками Парламентской библиотеки в сотрудничестве со специалистами различных организаций [14].

Как всякий информационно-поисковый тезаурус, созданный изначально для ручного индексирования, EUROVOC представляет собой искусственный язык, созданный на базе естественного языка предметной области. Специфика тезауруса, предназначенного для ручного индексирования, влечет за собой определенные проблемы при использовании его для автоматической обработки.

Прежде всего, представление дескрипторов тезауруса в тексте значительно более разнообразно, чем это указано в русской версии тезауруса EUROVOC. Например, дескриптор *ОХРАНА ОКРУЖАЮЩЕЙ СРЕДЫ* помимо указанных в тезаурусе вариантов может быть выражен также следующими словами и терминами, неописанными в тезаурусе, но встречающимися в текстах российских правовых актов: *защита природы, природозащитный, природоохранный, природоохранительный (меры, деятельность, процесс*).

В тезаурусах для ручного индексирования обычно не указывается неоднозначность некоторых терминов, описанных в русской версии только в одном из значений, что не нужно для человека-индексатора, но необходимо для автоматической обработки. Примеры неоднозначных терминов тезауруса, включенных в русскую версию EUROVOC в одном значении, что может привести к неправильному автоматическому индексированию: *кожа* (как кожевенная продукция и кожа человека), *печать* (как СМИ, как штамп, как процесс печатания), *питание* (еда и электрическое питание), *корма* (питание животных и часть корабля), *образование* (как обучение и как создание чего либо).

Средства описания и работы с многозначностью необходимы для любого ресурса, использующегося для автоматической обработки текстов.

Для тезаурусов, предназначенных для ручного индексирования характерно, что иерархия понятий останавливается на достаточно высоком уровне иерархии и не включает более конкретные термины. Между тем, например, среди нормативных документов широко представлены такие документы, в которых обсуждается *минтай*, но нет слова *рыба*, обсуждаются *солдаты,* но нет слова *военнослужащий*, обсуждается *пшеница*, но нет слова *зерно* и многие другие подобные примеры. Такие тексты не могут быть автоматически проиндексированы правильно из-за нехватки информации в тезаурусе EUROVOC.

Автоматическое индексирование предполагает и автоматизацию поиска, то есть поиск с автоматическим расширением запроса. В связи с этим большую значимость приобретает качество описания отношений между дескрипторами в тезаурусе. В тезаурусах для ручного индексирования обычно используются отношения ВЫШЕ-НИЖЕ и АССОЦИАЦИЯ. На широко представленные в тезаурусе EUROVOC отношения ассоциации невозможно уверенно опереться при автоматическом расширении запроса, например:

МОНОГРАФИИ

АСЦ *ТИПОГРАФИИ*

Ищем тексты о монографиях, получаем тексты о типографиях, среди которых не так много текстов посвящено публикации монографий, и наоборот.

Отметим, что эти проблемы возникают уже в одноязычной среде, а в многоязычной значительно усугубляются. В связи с этими проблемами тезаурусы, созданные для индексирования человеком-индексатором, невозможно использовать на практике в автоматическом режиме как в одноязычной, так и многоязычной среде.

3. Традиционные двуязычные словари и лингвистические ресурсы для автоматического концептуального индексирования

В предыдущем параграфе мы показали, что тезаурус для автоматического концептуального индексирования должен быть значительно подробнее, чем традиционные одноязычные и многоязычные информационно-поисковые

тезаурусы. По подробности словарного состава такой тезаурус близок к двуязычным терминологическим словарям, однако имеет некоторые серьезные особенности.

Прежде всего, двуязычный тезаурус предоставляет не пословные переводы, а организован на основе понятий: список языковых выражений на русском языке – название понятия на русском языке – название понятия на английском языке - список текстовых вариантов на английском языке. При этом списки языковых вариантов понятия должны быть как можно больше с каждой стороны, включая и те выражения, которые обычно не представлены в словарях, так как представляются очевидными для человека. Возможности автоматического процесса значительно скромнее, кроме того, наличие многословных вариантов снижает проблему разрешения лексической многозначности.

Так, например, в Общественно-политическом Тезаурусе понятие *ЗДРАВООХРАНЕНИЕ* имеет следующий набор вариантов на русском и английском языках:

Русские варианты: *защита здоровья, здравоохранительный, здравоохранительные меры, обеспечение здоровья, общественное здравоохранение, оздоровление граждан, оздоровление населения, охрана здоровья, система здравоохранения*

Английские варианты: *public health, community health, health care, health care sector, health care system, health field, health of population, health promotion, provision of health, public health field.*

Другой важной особенностью разработки тезауруса для автоматического концептуального индексирования является то, что все возможные текстовые варианты на обоих языках должны быть эквивалентны относительно тезаурусных связей, например, находиться на приблизительно одном уровне иерархии.

При создании традиционных словарей ситуация несколько иная. Основной целью традиционных двуязычных словарей является обеспечение совокупности наиболее частых переводов слова в различных текстах. Переводы даются как бы с запасом, в список переводов включаются и точные переводы, и переводы с более узким значением и с более широким (именно поэтому англо-русские и русско-английские словари не являются обратимыми). Предполагается, что читающий разберется по контексту, какой перевод выбрать.

Пока компьютерные системы не имеют достаточных аналитических способностей, чтобы в полной мере учитывать контекст. В то же время тезаурусная сеть позволяет наглядно представить соотношение понятий – более узкое представить нижестоящим понятием, более широкое - вышестоящим понятием.

Термин, который не имеет адекватного перевода, может остаться совсем без перевода, однако по тезаурусной сети можно легко выяснить ближайшие по смыслу более узкие и более широкие понятия. Часто оказывается, что точное значение того или иного понятия можно представить достаточно употребительным словосочетанием. Например, электронный словарь МультиЛекс 1.0a [16, 17] представляет значение существительного *sabotage* следующим образом:

sabotage I n

1. саботаж
2. диверсия; подрывная деятельность; вредительство

В русской части Общественно-политического тезауруса перечисленным словам соответствуют три различных понятия: *САБОТАЖ, ДИВЕРСИЯ, ВРЕДИТЕЛЬСТВО*. Если буквально приписать слово sabotage ко всем этим трем понятиям, то мы получим, что у слова три значения. Но это не соответствует англоязычным источникам, которые видят одно соответствующее значение:

Sabotage - any underhand interference with production, work, etc., in a plant, factory, etc., as by enemy agents during wartime or by employees during a trade dispute. *[8]*

Sabotage – any deliberate destruction, disruption, or damage of equipment, a public service, etc., as by enemy agents, dissatisfied employees, etc. *[3]*

Анализ употребления английского слова sabotage *и соответствующих слов русского языка приводит к необходимости установления отношений гипонимии между понятиями* ВРЕДИТЕЛЬСТВО - САБОТАЖ, ВРЕДИТЕЛЬСТВО – ДИВЕРСИЯ *(этих отношений до начала работы со словом* sabotage *не было – так сопоставительный анализ значений английских слов приводит к более качественному описанию значений русских слов). Понятие* SABOTAGE *было поставлено в соответствие понятию* ВРЕДИТЕЛЬСТВО. *Англоязычный ряд для русского понятия* САБОТАЖ *имеет вид списка словосочетаний:* employee sabotage, labor sabotage, sabotage by employees, silent sabotage, workers sabotage. *Англоязычный ряд для русского понятия* ДИВЕРСИЯ *таков:* sabotage attack, enemy sabotage, sabotage by enemy, sabotage explosion.

В качестве другого примера рассмотрим значение англоязычного слова brother-in-law *(брат супруга/супруги), для которого в русском языке нет ни соответствующего слова, ни единого употребительного словосочетания. В таких случаях заводится понятие со специальной пометкой «#», обозначающей, что русского эквивалента нет. Понятие снабжается русским пояснением. Отношение с другими понятиями тезауруса показывает соотношение русских и английских понятий:*

BROTHER-IN-LAW - # ДЕВЕРЬ ИЛИ ШУРИН

ВЫШЕ KINSMAN - РОДСТВЕННИК-МУЖЧИНА

НИЖЕ BROTHER OF HUSBAND - ДЕВЕРЬ

НИЖЕ BROTHER OF WIFE - ШУРИН

Таким образом, основными особенностями двуязычного тезауруса, предназначенного для автоматического концептуального индексирования являются:

- описание как можно более точных соответствий между терминами путем размещения их в синонимические ряды одного и того же понятия, в том числе оставление термина без перевода, если этого перевода нет;
- описание как можно большего количества синонимических вариантов выражения понятия в тексте для обоих языков, что является базой для распознавания понятия в тексте;
- описание многозначности терминов обоих языков;
- описание как можно большего количества многословных синонимических вариантов, что облегчает процедуру разрешения неоднозначности терминов в тексте.

4. Основные этапы разработки двуязычного тезауруса

На первом этапе русскоязычные термины тезауруса были переведены на английский язык с помощью традиционных словарей. Было получено 33 тысячи англоязычных терминов. Однако при этом вне тезауруса остались термины, свойственные общественно- политической жизни англоязычных стран и не присутствующие в России, кроме того синонимическая вариативность англоязычной части была представлена недостаточно широко.

Поэтому на следующем этапе был взят комплекс наиболее известных английских и американских толковых словарей, терминологических словарей по бизнесу и праву, информационно-поисковых тезаурусов и организована вычитка этих словарных источников на предмет выявления новых синонимов уже описанных понятий и новых важных понятий. При этом термины могли извлекаться не только из заголовков словарных статей, но часто из примеров и толкований. Эта часть практически завершена, что привело к описанию в тезаурусе 55 тысяч англоязычных терминов.

На следующем этапе осуществляется вычитка и корректировка англоязычных синонимичных рядов тезауруса, при этом осуществляется контроль по доступным в Интернет текстам подавляющего числа накопленных переводных терминов. Кроме этого, эксперт может предложить дополнительные варианты синонимичных многословных терминов и проверить по Интернет их реальное существование. Выяснилось, что традиционные словарные источники дают множество неупотребительных конструкций или конструкций, поменявших свое значение или имеющих и другие значения, кроме указанных.

На текущем этапе параллельно начата тестовая автоматическая обработка больших массивов англоязычных текстов, включающая автоматическое концептуальное индексирование, рубрицирование и аннотирование текстов.

Представляется, что наиболее эффективным средством для выявления неточностей и неполноты представления англоязычной терминологии является процесс автоматической рубрикации, например, по Классификатору правовых актов (1069 рубрик) коллекций официальных англоязычных документов. Мы реализовали эту систему автоматической рубрикации в 2001 году, в 2002 году она прошла тестирование на 10 тысячах документах федерального уровня и была применена для автоматической рубрикации более 160 тысяч документов регионального уровня [12]. Особенностью системы является то, что она базируется на построенном для документа концептуальном индексе и поэтому не зависит от языка документа.

Имеющаяся программная оболочка позволяет:

- выполнить рубрикацию заданного набора текстов,
- для каждой рубрики, проставленной конкретному тексту, просмотреть набор терминов, на основе которых была получена данная рубрика,

- для каждого термина в тексте увидеть его интерпретацию в виде понятия тезауруса и результаты процесса разрешения многозначности.

Таким образом, англоязычные тексты будут автоматически рубрицироваться. Эксперты будут просматривать документы, оценивать основное содержание и сравнивать с набором полученных рубрик. Появление «странных» (с точки зрения эксперта) рубрик обычно свидетельствует о неточном описании терминов, нехватка рубрик – о недостаточной полноте описания. Такой процесс позволит сконцентрироваться на возможных неточностях в иерархии тезаурусных описаний и даст возможность планомерно их устранять.

5. Поиск двуязычных документов в УИС РОССИЯ

Для документов на русском или английском языке производится их тематический анализ. В результате тематического анализа для документа создается концептуальный индекс, независящий об исходного языка документа и конкретных синонимов, употребленных в тексте. Каждое понятие в концептуальном индексе имеет вес, построенный как на основе частотных характеристик употребления понятий в тексте, так и его тезаурусных и текстовых связей с другими понятиями [11].

Помимо стандартных поисковых атрибутов, включающих средства контекстного поиска, использующие результаты морфологической обработки русских, английских и смешанных русско-английских текстов, в УИС РОССИЯ разработаны как русскоязычный, так и англоязычный интерфейс тематического поиска.

Возможен поиск с опцией «расширение по дереву», когда релевантными считаются документы, содержащие не только синонимы выбранного понятия, но синонимы подчиненных понятий. В многоязычной коллекции результатом поиска могут быть ссылки на документы на разных языках (см. Рис. 1).

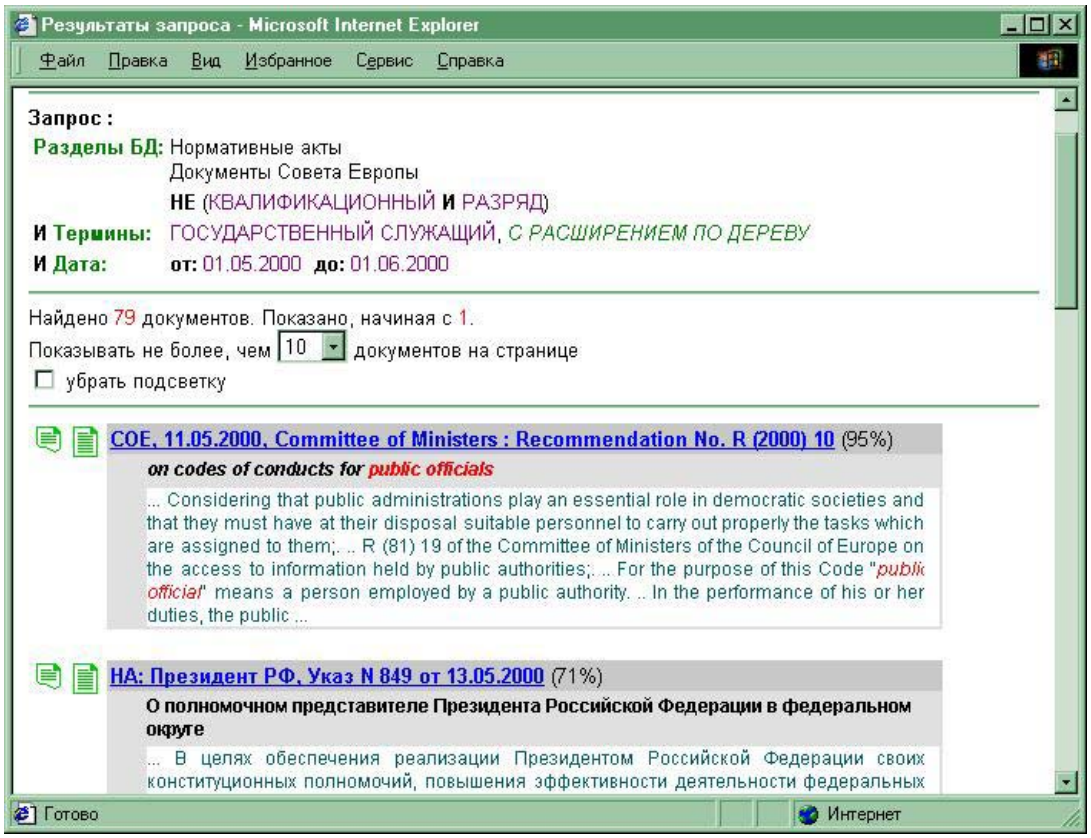

Рис.1. Результаты поиска в многоязычной коллекции

При тематическом поиске важное значение приобретает обоснование релевантности найденного документа, особенно для документов на другом языке. В УИС РОССИЯ релевантные запросу вхождения терминов подсвечиваются в тексте документа (см. Рис.2).

Отметим, что в УИС РОССИЯ реализованы средства интерактивного уточнения запроса путем отображения на специальной панели наиболее «важных» понятий в результатах выполненных пользователем запросов. Одним нажатием кнопки «мыши» пользователь может добавить новое условие в запрос (AND, AND NOT, в том числе с

опцией «расширение по дереву»). В двуязычной среде эти средства приобретают новое качество – пользователь может формировать нужную ему подборку документов, работая до последнего момента в родной языковой среде.

Рис.2. Обоснование результатов поиска

Заключение

Мы описали основные принципы организации, этапы построения и тестирования, а также способы использования двуязычного русско-английского Тезауруса по общественно-политической жизни, создаваемого как лингвистический ресурс для автоматического концептуального индексирования и тематического поиска русскоязычных и англоязычных текстов.

Благодарности

Эта работа частично выполняется при поддержке гранта #01-07-430 Российского фонда фундаментальных исследований.

Литература

1. Braschler M., CLEF 2001 - Overview of Results // Evaluation of Cross-Language Information Retrieval Systems / C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.) - Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001. Revised Papers – Lecture Notes in Computer Science, 2406 - Springer-Verlag: Berlin, Heidelberg, 2002.
2. (http://link.springer.de/link/service/series/0558/papers/2406/24060009.pdf)
3. Climent S., Rodriguez H., Gonzalo J., Definition of the link and subsets for nouns of the EuroWordNet project - EuroWordNet (LE2-4003), Technical Report D005.
4. (http://sensei.ieec.uned.es/~julio/D005.ps)
5. Collins Electronic Dictionary & Thesaurus – v.1.5 - London: Harper-Collins, 1995.
6. Gay F.C., Oard D.W., The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French or Arabic Queries // NIST Special Publication 500-250: The Tenth Text Retrieval Conference (TREC-2001). (http://trec.nist.gov/pubs/trec10/papers/clirtrack.pdf)

7. Gonzalo J., Language Resources in Cross-Language Information Retrieval: a CLEF perspective // Cross-Language Information Retrieval and Evaluation: Proceedings of the First Cross-Language Evaluation Forum, LNCS, Springer-Verlag.
8. (http://sensei.lsi.uned.es/NLP/papers/clef00.pdf)
9. Kando N., Kuriyama K., Yoshioka M., Overview of Japanese and English Imformation Retrieval Tasks (JEIR) at the Second NTCIR Workshop // NTCIR Workshop 2. Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization May 2000 - March 2001. - National Institute of Informatics, Tokyo, Japan – 2001.
10. (http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/ovview-kando2.pdf)
11. Mauldin M.L., Retrieval Performance in FERRET: A Conceptual Information Retrieval System The 14th International Conference on Research and Development in Information Retrieval. – 1991.
12. Random House Webster's Unabridged Dictionary (Version 3.0 for Windows 95/98/NT & Version 2.2 for Windows 3.1).
13. Richardson R., Smeaton A., Using WordNet in a Knowledge-Based Approach to Information Retrieval. School of Computer Applications. Working Paper, CA-0395. 1995.
14. Woods W.A., Conceptual indexing: a better way to organize knowledge. SunMicrosystems Laboratories Technical Report, SMLI TR-97-61. - 1997.
15. Добров Б.В., Лукашевич Н.В., Тезаурус и автоматическое концептуальное индексирование в Университетской Информационной Системе РОССИЯ // Сборник трудов Третьей Всероссийской конференции по Электронным Библиотекам: "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" (RCDL'2001) - Петрозаводск, 2001. - С.78-82.
16. Добров Б.В., Лукашевич Н.В., Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту. КИИ-2002. 7-12 октября 2002, Коломна – М.: Физматлит – Т.1 – С.178-186.
17. Журавлев С.В., Юдина Т.Н., Информационная система РОССИЯ // НТИ, Сер.2. 1995. - № 3. С. 18-20.
18. Информационно-поисковый тезаурус. Русская версия тезауруса EUROVOC. В 3х томах. - Издание Государственной Думы ФС РФ, 2001.
19. Лукашевич Н.В., Салий А.Д., Представление знаний в системе автоматической обработки текстов // НТИ, Сер.2. 1997. № 3. С. 1-6.
20. МультиЛекс 1.0а. – Англо-русский электронный словарь – ЗАО «МедиаЛингва», 1996.
21. Новый большой англо-русский словарь. – В 3х томах. – Под ред. Ю.Д.Апресяна, Е.М.Медниковой – М.: Русский язык, 1993-94.