АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

Г. Г. Белоногов, Р. С. Гиляревский, М. В. Козачук,
Ал-др А. Хорошилов, Ал-сей А. Хорошилов

О возможности поиска информации в русскоязычных базах данных ВИНИТИ по запросам, сформулированным на основных европейских языках

Рассматриваются принципы построения системы поиска информации в русскоязычных базах данных по запросам, сформулированным на основных европейских языках, с выдачей результатов поиска на английском и на русском языках. Работа проводится при финансовой поддержке Российского Фонда Фундаментальных Исследований.

ВИНИТИ РАН является крупнейшим в мире центром научно-технической информации, издающим реферативные журналы по естественным и техническим наукам. Наряду с реферативными журналами здесь выпускаются также базы данных на машиночитаемых носителях. В результате этой деятельности в ВИНИТИ создан ретрофонд, содержащий более двадцати миллионов описаний публикаций отечественных и зарубежных авторов. Ретрофонд ВИНИТИ является важным национальным информационным ресурсом, отражающим научные и научно-технические достижения второй половины XX в. К сожалению, из-за языковых барьеров, доступ к этому ценнейшему источнику знаний существенно ограничен.

Проблему доступа к информации можно было бы решить путем создания системы поиска в реферативных базах данных по запросам, сформулированным на различных "естественных" языках, что облегчило бы работу не только зарубежным, но и отечественным ученым и специалистам, так как тогда отпала бы необходимость в знании правил формализации запросов.

Первым шагом в этом направлении явилось создание в 2000 г. и установка на сервере ВИНИТИ системы поиска информации в русскоязычных базах данных по неформализованным запросам, сформулированным на английском языке, с выдачей результатов поиска также на английском языке [1]. Система функционирует под управлением операционной системы UNIX. Обращаться к ней можно как в рамках сети Internet, так и в рамках локальной сети ВИНИТИ.

Быстрое создание системы доступа к базам данных ВИНИТИ по запросам на английском языке оказалось возможным только благодаря тому, что в ее основу были положены системы русско-английского и англо-русского фразеологического машинного перевода текстов (системы RETRANS и ERTRANS [2]), разработанные ранее коллективом ученых под руководством профессора Г. Г. Белоногова. На это ушло более двенадцати лет. А как быть с другими языками? Ведь нельзя же всерьез думать, что за какие-нибудь два года (срок выполнения проекта РФФИ) можно ограниченными силами разработать системы машинного перевода для таких пар европейских языков, как русский и немецкий, русский и французский, русский и испанский! Но выход все-таки есть.

Дело в том, что в современном мире (так уж случилось) английский язык все более и более становится языком межнационального общения, и квалифицированные ученые и специалисты европейских стран, как правило, в той или иной степени владеют этим языком. Причем им обычно бывает легче понять текст, представленный на английском языке, чем сформулировать поисковый запрос на этом языке.

В этой связи напрашивается следующее решение проблемы: создать такую систему поиска информации в русскоязычных базах данных, в которой поисковые запросы будут формулироваться на родном языке (например, на русском, английском, немецком, французском или испанском), а результаты поиска — выдаваться на английском или (для русскоязычных пользователей) на русском языке. Тогда при подключении к системе каждого нового языка запросов достаточно разработать только систему автоматического перевода на русский язык и автоматической формализации запросов.

Система поиска информации в русскоязычных базах данных по запросам на английском языке, созданная в ВИНИТИ, состоит из четырех компонент: 1) реферативной базы данных объемом более 20 млн записей; 2) СУБД типа "Сокол"; 3) системы автоматического перевода на русский язык и автоматической формализации запросов, представленных на английском языке; 4) системы автоматического перевода результатов поиска информации с русского языка на английский. При такой структуре системы достаточно для каждого нового входного языка создать компоненту, аналогичную компоненте № 3, и можно будет обращаться к системе с запросами на этом языке.

В 2001 г. в рамках проекта РФФИ были разработаны принципы автоматического перевода и автоматической формализации поисковых запросов, представленных на русском, немецком, французском и испанском языках. Это создало предпосылки для дальнейшего развития системы мультиязычного доступа к базам данных ВИНИТИ. Рассмотрим эти принципы.

Автоматический перевод на русский язык и автоматическая формализация поисковых запросов, сформулированных на иностранных языках, осуществляются в следующем порядке. Сначала выполняется морфологический анализ текста запроса. Затем — его семантико-синтаксический и концептуальный анализ, в результате которого текст запроса представляется в виде ряда наименований понятий, которым ставятся в соответствие их переводные эквиваленты на русском языке, их синонимы и квазисинонимы. Далее производится морфологический анализ русских переводных эквивалентов иноязычных наименований понятий, и по результатам этого анализа из их состава исключаются малоинформативные лексические единицы: местоимения, предлоги, союзы, частицы и глаголы (опыт эксплуатации поисковых систем показал, что при тематическом поиске эти лексические единицы играют незначительную роль).

Формализация запросов заключается в проставлении между их информативными словами логических связок и так называемых "синтаксических". операторов, характеризующих вхождение этих слов в одно словосочетание, в одно предложение, в один параграф и в один документ. Приоритетность логических связок и "синтаксических" операторов определяется их характером и системой скобок. Отождествление различных словоизменительных форм русских слов в процессе поиска информации осуществляется путем усечения этих слов на основе результатов их автоматического морфологического анализа.

Повышение полноты поиска информации достигается за счет использования синонимических отношений между русскоязычными переводными эквивалентами иноязычных слов и словосочетаний и специального словаря русских синонимов и гипонимов, составленного на основе обработки более 70-ти тезаурусов, созданных в рамках ГАСНТИ СССР. Более подробно принципы автоматического перевода на русский язык и автоматической формализация поисковых запросов, представленных на иностранных языках, описаны на примере англоязычных запросов в статье [2].

Автоматическая формализация русскоязычных запросов, обращенных к русскоязычным базам данных, ведется по существу по тем же правилам, что и автоматический перевод на русский язык и формализация поисковых запросов, представленных на иностранных языках. Разница состоит лишь в том, что вместо этапа перевода запроса с иностранного языка на русский здесь вводится этап автоматического семантико-синтаксического и концептуального анализа русскоязычного запроса — его расчленения на наименования понятий.

Это делается с помощью того же комплекса программ, который реализует этап семантико-синтаксического и концептуального анализа при переводе текстов с русского языка на английский. Только вместо русско-английских словарей здесь используются "русско-русские" словари, в которых русским словам и словосочетаниям ставятся в соответствие те же самые слова и словосочетания.

Сложнее дело обстоит с автоматическим переводом на русский язык и формализацией запросов, представленных на языках, отличных от английского и русского (например, на немецком, французском и испанском языках). Здесь уже не удастся ограничиться только процедурными и словарными средствами системы машинного перевода RETRANS. Для этих языков придется, как минимум, разрабатывать соответствующие процедуры морфологического анализа и создавать немецко-русские, франко-русские и испано-русские машинные словари. Задача эта не из простых и потребует немало времени и сил.

Рассмотрим сначала возможности создания процедур морфологического анализа. Традиционный подход здесь потребовал бы выполнения следующих работ: 1) составления словника большого объема путем компиляции словников из ранее составленных словарей или, что значительно лучше, путем автоматического составления словника по текстам большого объема; 2) назначения каждому элементу словника соответствующей ему грамматической информации (части речи, длины основы, типа словоизменения и др.); 3) составления грамматических таблиц для морфологического анализа; 4) разработки алгоритма и программы морфологического анализа.

Традиционный подход хорош тем, что создаваемые на его основе процедуры обеспечивают высокую точность морфологического анализа. Но здесь возникает проблема "новых" слов — слов, не включенных в словарь. При таком подходе "новые" слова не будут распознаваться, и им не будет назначаться грамматическая информация. А это неизбежно отрицательно скажется на качестве автоматической обработки информации.

Авторы статьи пошли по другому пути — по пути создания процедур морфологического анализа на основе применения метода аналогии. Этот метод базируется на том факте, что в основных европейских языках (русском, немецком, английском, французском, испанском, итальянском, чешском, словенском, сербском, польском, румынском и др.) имеет место сильная корреляция между буквенным составом концов слов (не обязательно суффиксов и окончания) и грамматической информацией к этим словам.

В таблице приведено несколько фрагментов обратного словаря слов, составленного по немецким текстам. В этих фрагментах указаны длины окончаний слов (первая слева цифра) и индексы обобщенной грамматической информации к словам (после косой черты).

Из таблицы видно, что значительные участки обратного словаря имеют одинаковые индексы обобщенной грамматической информации и одинаковую длину окончаний. Это позволяет при построении процедуры морфологического анализа с использованием метода аналогии построить компактную таблицу, позволяющую определять грамматические характеристики слов по их конечным буквосочетаниям. При этом характеристики слов, входивших в состав исходного словаря, по которому строилась таблица, будут определяться правильно с вероятностью 100%, а всех прочих слов немецкого языка — с высокой вероятностью (98–99%). Аналогичная картина наблюдается также у обратных словарей французского и испанского языков. Более подробно с принципами построения алгоритмов морфологического анализа с использованием метода аналогии можно ознакомиться по статьям [3, 4].

Фрагменты обратного словаря немецких слов

Panorama 1/11
Thema 1/11
Schema 1/11
Paradigma 1/11
Asthma 1/11
Klima 1/11
Dilemma 1/11
Komma 1/11
Firma 1/11
.....
Band 0/11
Hand 0/11
Land 0/11
Rand 0/11
Sand 0/11
Wand 0/11
.....
bebend 0/25
gebend 0/25
nachgebend 0/25
massgebend 0/25
hebend 0/25
erhebend 0/25
schiebend 0/25
lebend 0/25
belebend 0/25
schwebend 0/25
reibend 0/25
lobend 0/25
sterbend 0/25
schnaubend 0/25
zeitraubend 0/25
zerstaeubend 0/25
badend 0/25
.....
lebende 1/25
belebende 1/25
strebende 1/25
widerstrebende 1/25
emporstrebende 1/25
betruebende 1/25
schwebende 1/25
bleibende 1/25
beschreibende 1/25
sterbende 1/25
zeitraubende 1/25
einladende 1/25
.....
wichtige 1/21
gewichtige 1/21
unwichtige 1/21
seitige 1/21
gegenseitige 1/21
jenseitige 1/21
diesseitige 1/21
zeitige 1/21
gleichzeitige 1/21
kurzzeitige 1/21
faltige 1/21
nachhaltige 1/21
.....
Heile 1/14
Meile 1/11
Seile 1/11
Teile 1/11
Weile 1/11
Zeile 1/11
.....
stampfte 1/33
kaempfte 1/33
niederkaempfte 1/33
schimpfte 1/33
klopfte 1/33
tropfte 1/33
stopfte 1/33
bezopfte 1/33
zupfte 1/33
schaerfte 1/33
einschaerfte 1/33
duerfte 1/34
beduerfte 1/34
.....
wahrhaftig 0/21
geschaeftig 0/21
kraeftig 0/21
heftig 0/21
giftig 0/21
triftig 0/21
vernuenftig 0/21
beduerftig 0/21
notduerftig 0/21
bedaechtig 0/21
verdaechtig 0/21
maechtig 0/21
.....
Ausbildung 0/14
Vergoldung 0/14
Befremdung 0/14
Verleumdung 0/14
Landung 0/14
Brandung 0/14
Sendung 0/14
Wendung 0/14
Vollendung 0/14
Muendung 0/14
.....
herzlich 0/21
schmerzlich 0/21
kuerzlich 0/21
grundsaetzlich 0/21
zusaetzlich 0/21
unverletzlich 0/21
ergoetzlich 0/21
ploetzlich 0/21
gesetzlich 0/21
.....
daemmern 0/32
haemmern 0/32
kuemmern 0/32
bekuemmern 0/32
verkuemmern 0/32
.....
liebender 2/25
lebender 2/25
sterbender 2/25
betaeubender 2/25
einladender 2/25
blendender 2/25
schlafender 2/25
.....
eigentliches 2/21
zaertliches 2/21
veantwortliches 2/21
koestliches 2/21
aengstliches 2/21
christliches 2/21
stattliches 2/21
unerbittliches 2/21
deutliches 2/21
trauliches 2/21
herzliches 2/21
schmerzliches 2/21
.....
Raritaet 0/11
Popularitaet 0/11
Integritaet 0/11
Prioritaet 0/11
Autoritaet 0/11
Virtuositaet 0/11
Universitaet 0/11
Quantitaet 0/11
Identitaet 0/11
Passivitaet 0/11
Aggressivitaet 0/11
Aktivitaet 0/11
.....
widerlegt 1/32
hinterlegt 1/32
verlegt 1/32
zerlegt 1/32
.....
Klarheit 0/11
Besonderheit 0/11
Sicherheit 0/11
Treffsicherheit 0/11
Unsicherheit 0/11
Datensicherheit 0/11
Netzsicherheit 0/11
Duesterheit 0/11
Wahrheit 0/11
Unwahrheit 0/11
Mehrheit 0/11
.....
bekennt 1/32
erkennt 1/32
nennt 1/32
benennt 1/32
goennt 1/32
.....
zeigest 3/34
belaestigest 3/34
folgest 3/34
verlangest 3/34
singest 3/34
beherbergest 3/34
verbeugest 3/34
.....
heisst 1/32
reisst 1/32
niederreisst 1/32
zerreisst 1/32
ausreisst 1/32
weisst 1/32
vergisst 1/32
.....
Ersatz 0/14
Vorsatz 0/14
Hauptsatz 0/14
Fortsatz 0/14
Zusatz 0/14

Наиболее трудоемкими являются работы по составлению и ведению двуязычных машинных словарей наименований понятий, необходимых для концептуальных текстов. В настоящее время авторы статьи располагают англо-русским и русско-английским машинными словарями такого рода объемом более 1 млн 500 тыс. словарных статей каждый. Эти словари используются при переводе на русский язык и формализации англоязычных запросов, а также при переводе результатов поиска информации с русского языка на английский. Машинные словари для автоматического перевода и формализации запросов, поступающих на вход ИПС на немецком, французском и испанском языках, еще предстоит создать. На первом этапе работ эти словари будут составляться на основе многоязычного словаря ВНИИКИ Госстандарта объемом более 160 тыс. словарных статей. В дальнейшем для этой цели будут использоваться двуязычные заголовки документов, содержащиеся в базах данных ВИНИТИ, и словари в книжной форме.

Как уже было указано, система автоматического перевода на русский язык и автоматической формализации запросов, представленных на английском языке, а также система автоматического перевода результатов поиска информации с русского языка на английский уже созданы. В дальнейшем они будут лишь совершенствоваться. Для доступа к базам данных ВИНИТИ по запросам, сформулированным на других языках (например, на немецком, французском или испанском), дополнительно предстоит создать только средства их автоматического перевода на русский язык и автоматической формализации. Общая схема функционирования этих средств будет следующая: 1) морфологический анализ текста запроса; 2) семантико-синтаксический и концептуальный анализ запроса; 3) поиск по словарям русских переводных эквивалентов наименований поня-