АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 005:2

Белоногов Г. Г., Гиляревский Р. С., Егоров В. С.,
Новоселов А. П., Хорошилов А. А., Шогин А. Н.

Автоматический перевод на русский язык англоязычных запросов и их формализация при поиске информации в русскоязычных реферативных базах данных

Рассматривается проблема автоматизированного поиска информации в русскоязычных базах данных по запросам на английском языке. Описывается разработанная авторами система автоматического перевода запросов на русский язык и их автоматической формализации.

Автоматический поиск информации в текстах по запросам на "естественном языке" — давнишняя мечта многих разработчиков поисковых систем. Не вдаваясь в детали этой проблемы, можно сказать, что в полном объеме эта мечта будет осуществлена не скоро. Ведь речь здесь идет ни много ни мало как об автоматическом распознавании "смысла" запросов и о последующем сопоставлении этого "смысла" со "смыслом" текстов, в которых ведется поиск. А средства выражения этого "смысла" весьма многообразны: здесь и многообразие словоизменительных и словообразовательных форм слов, и явление лексической полисемии, синонимии и гипонимии, и синтаксическая синонимия, и явление эллипсиса, и многое другое.

Несколько проще дело обстоит с поиском информации в библиографических базах данных по тематическим запросам. Здесь за несколько десятилетий уже накопился некоторый опыт и сложились определенные традиции. Например, утвердилась традиция надстраивать над массивами реферативных баз данных некоторый поисковый аппарат в виде так называемого "инверсного файла" — словаря словоформ, в котором для каждой словоформы указываются все адреса ее вхождения в тексты. Поиск информации ведется по запросам, которые представляют собой последовательности слов, соединенных логическими связками AND, OR и NOT. Этим связкам соответствуют теоретико-множественные операции пересечения, объединения и вычитания множеств. Часто используются также так называемые "контекстуальные" операторы— операторы SAME, WITH и ADJ.

"Контекстуальные" операторы — это операторы типа AND, на выполнение которых накладываются дополнительные ограничения. Так, в случае оператора SAME требуется, чтобы наряду с выполнением условия одновременного вхождения в документ двух слов A и B они также входили и в одно и то же "поле" описания этого документа (например, в поле текста реферата или в поле заголовка документа). Это условие формулируется как "A SAME B". В случае оператора WITH требуется, чтобы слова A и B входили в одно и то же предложение (условие формулируется как "A WITH B"), а в случае оператора ADJ требуется выполнение условия контактного расположения слов A и B или их расположение на расстоянии одного или нескольких других слов (условие формулируется как "A ADJ B" или "A ADJ1 B" или "A ADJ2 B" и т. д.). При наличии в поисковом запросе нескольких логических и "синтаксических" операторов порядок их выполнения определяется их приоритетами и скобками. Операторы ADJ, WITH, SAME и AND имеют более высокий приоритет, чем оператор OR.

При формулировке запросов с использованием перечисленных операторов приходится прежде всего считаться с многообразием форм слов в текстах рефератов и заголовках документов. Ведь в процессе поиска информации необходимо обеспечить отождествление слов запросов и документов несмотря на различие их грамматических форм. Это можно сделать тремя способами: 1) путем лемматизации (приведения к основной словарной форме) всех слов запросов и слов инверсного файла; 2) путем генерации для слов запросов всех их словоизменительных и словообразовательных форм; 3) путем усечения слов запросов.

Первый способ нам представляется наиболее предпочтительным, и авторы статьи уже имели опыт его применения. Но при этом потребуется переформировать и перезагрузить в ЭВМ весь ранее накопленный массив баз данных, что потребует немалых трудозатрат. Применение второго способа приведет к резкому увеличению объема запроса (для русского языка, в случае генерации одних только словоизменительных форм, — в восемь раз). При третьем способе возникает опасность увеличения поискового "шума". Но если применять усечение слов только на границах их основ и окончаний, то уровень "шума" может оказаться вполне приемлемым.

При формулировке запросов желательно так-

же воздерживаться от использования малоинформативных (например, служебных) слов и, по возможности, позаботиться о включении в запросы синонимов и гипонимов (более узких по смыслу информативных слов).

В последние годы, в связи с развитием глобальной информационной сети Internet, существенно возросла актуальность решения проблемы преодоления языковых барьеров при поиске информации в разноязычных базах данных по запросам на "естественном" (желательно родном) языке. Причиной этому являются, как минимум, два фактора: слабое знание большинством пользователей иностранных языков и слабое знание правил формализации запросов. Проблема эта весьма сложная, так как ее решение должно опираться на решение проблемы машинного перевода текстов с одних естественных языков на другие и проблемы автоматической формализации поисковых запросов (их перевода с "естественного" языка на формализованный "информационный" язык).

В ВИНИТИ в течение последних десяти-пятнадцати лет велись исследования и разработки, направленные на создание промышленных систем русско-английского (RETRANS) и англо-русского (ERTRANS) машинного перевода текстов. В основу этих исследований и разработок была положена концепция фразеологического машинного перевода [1, 2]. В настоящее время обе эти системы установлены на сервере ВИНИТИ и работают под управлением операционной системы UNIX. Каждая из систем включает в свой состав политематический словарь объемом более полутора миллионов лексических единиц (преимущественно словосочетаний) и двенадцать настроечных тематических словарей общим объемом более 200 тыс. словарных статей. Системы RETRANS и ERTRANS реализованы также в среде операционных систем MS DOS, WINDOWS 95, WINDOWS 98 и WINDOWS NT.

В течение 1993–1994 гг. на основе систем RETRANS и ERTRANS под руководством Г. Г. Белоногова и Б. А. Кузнецова была предпринята успешная попытка создания системы поиска информации в русскоязычных базах данных по запросам на английском языке с выдачей результатов поиска также на английском языке (система BROWSER). Эта система могла работать на IBM-совместимых персональных компьютерах в среде MS DOS.

В 1999 г. на основе более мощных версий систем RETRANS и ERTRANS авторами статьи была создана другая система поиска информации в русскоязычных базах данных по формализованным запросам на английском языке. Она была установлена на сервере ВИНИТИ и могла работать в среде операционной системы UNIX в режиме массового обслуживания.

Обращение к системе с формализованными запросами связано с двумя недостатками: 1) от ее пользователя требуется знание правил формализации; 2) формализованные запросы переводятся на русский язык пословно и преимущества фразеологического перевода текстов никак не проявляются. Поэтому было принято решение создать систему поиска информации по неформализованным запросам с их автоматическим переводом на русский язык и последующей автоматической формализацией.

Перевод и формализация англоязычного запроса осуществляется в следующем порядке. Сначала проводится его семантико-синтаксический и концептуальный анализ. В результате анализа из текста запроса вычленяются наименования понятий, которым ставятся в соответствие русские переводные эквиваленты или последовательности переводных эквивалентов. Затем по результатам концептуального анализа составляется англо-русский словарь, и на его базе выполняется первый этап формализации запроса.

При выполнении первого этапа формализации можно исходить из следующих соображений:

1. Все понятия, отраженные в запросе, должны одновременно содержаться в искомых документах. Поэтому соответствующие им русские переводные эквиваленты или последовательности переводных эквивалентов должны соединяться друг с другом знаком AND.

2. Если одному английскому наименованию понятия ставятся в соответствие несколько русских переводных эквивалентов, то эти переводные эквиваленты должны соединяться друг с другом знаком OR. То же самое следует делать и в тех случаях, когда одному английскому слову ставятся в соответствии несколько словообразовательных вариантов его переводных эквивалентов.

3. Если переводной эквивалент английского наименования понятия выражается русским словосочетанием, то слова, входящие в состав этого словосочетания, должны соединяться друг с другом знаком ADJ.

4. Русские служебные слова, местоимения и глаголы должны исключаться из запроса как малоинформативные элементы. Это можно делать, опираясь на результаты морфологического анализа. По результатам морфологического анализа можно также производить усечение слов. Оно проводится по границе между основой и окончанием слова. После знака усечения должна ставиться цифра "3" или "5" (цифра "3" — обозначает максимальную длину окончания слова при отсутствии у него возрастной частицы, а цифра "5" — при ее наличии). Если в процессе поиска информации начальная часть слова текста совпадает с основой слова из запроса, а число букв в оставшейся части слова не превосходит число, стоящее после знака усечения, то слово из текста и слово из запроса считаются тождественными по смыслу. Если из словосочетания исключается какой-либо малоинформативный элемент, не стоящий в его начале или в конце, то вместо этого слова и двух связок ADJ слева и справа от него ставится связка ADJ1.

Соображения, предлагаемые в качестве основы при выполнении первого этапа формализации запроса, в определенной мере отражают опыт построения поисковых систем, но имеют, как минимум, два существующих недостатка: 1) при большом числе наименований понятий, выделяемых в процессе концептуального анализа запроса, мала вероятность их одновременного вхождения в тексты рефератов документов, 2) при малом числе таких наименований велика вероятность возникновения поискового шума.

Эти недостатки можно устранить путем одновременного использования различных способов формализации запроса и введения нескольких эшелонов выдачи информации. При этом для первого эшелона могут быть сформулированы наиболее

жесткие логические условия, для второго — послабее, а для последующих эшелонов — еще слабее. Но если принимается решение иметь только один эшелон выдачи информации, то придется идти на компромисс.

Одно из возможных компромиссных решений при выполнении второго (уточняющего) этапа формализации запроса может быть следующим:

1) запросы, состоящие из одного концепта (словосочетания или слова), не модифицируются;

2) модификация запросов, состоящих из двух концептов, должна заключаться в замене связки AND между концептами на связку WITH, выражающую условие одновременного вхождения двух концептов в одно предложение;

3) при модификации запросов, состоящих из трех концептов (например, из концептов A, B и C), соседние концепты попарно соединяются связкой WITH, а между парами концептов ставится связка OR. Логическая формула запроса будет иметь вид:

A WITH B OR B WITH C;

4) при модификации запросов, состоящих более чем из трех элементов, сначала все они попарно соединяются связкой WITH, и между парами концептов ставится связка OR. Далее первые две пары концептов и отдельно все последующие их пары заключаются в скобки, а между скобками ставится связка SAME, выражающая условие вхождения в одно поле документа. Например, для запроса, состоящего из шести концептов A, B, C, D, E и F, логическая формула будет иметь вид:

(A WITH B OR B WITH C) SAME (C WITH D OR D WITH E OR E WITH F)

При выборе правил второго этапа формализации запросов мы учитывали такое явление как полисемия слов, которая выражается в том, что английским словам и, значительно реже, словосочетаниям могут быть поставлены в соответствие перечни различных по смыслу русских переводных эквивалентов. Эта полисемия может быть разрешена с помощью контекста. В качестве контекста могут выступать тексты рефератов и заголовки библиографических баз данных, если потребовать, чтобы переводные эквиваленты двух рядов расположенных слов и словосочетаний запроса входили в одно и то же предложение реферата или заголовка документа, т. е. чтобы выполнялось логическое условие WITH. При этом будут "работать" только осмысленные (по данному контексту) сочетания переводных эквивалентов.

Другое важное соображение. Если вероятность одновременного вхождения в тексты рефератов всех концептов запроса мала, и требуется пожертвовать частью их, то какой именно? В связи с этим уместно напоминать, что поисковые запросы представляют собой, как правило, номинативные предложения, в которых основная смысловая нагрузка приходится на их начальную часть. Поэтому разумно считать, что вхождение в поисковый массив концептов начальной части исходного запроса более важно, чем всех других его концептов. Другие концепты выражают, как правило, дополнительные поисковые признаки, уточняющие признаки, содержащиеся в левой части запроса.

Исходя из описанных принципов авторы статьи построили и установили на сервере ВИНИТИ комплекс программ автоматического перевода на русский язык и формализации англоязычных запросов с целью поиска информации в русскоязычных реферативных базах данных. В табл. 1 представлен пример перевода и формализации запроса.

Таблица 1

Перевод и формализация запроса
Production technique of half-finished goods of composites material

a) Концептуальный анализ запроса

Production	00001	00002	технология производства / способ изготовления / способы получения / способы производства
technique	00002		
of	00003		
half-finished	00004	00005	полуфабрикаты
goods	00005		
of	00006		
composites	00007	00008	композиционные материалы
material	00008		
•	00009		

b) Англо-русский словарь наименований понятий

Production technique * технология производства / способ изготовления / способы получения / способы производства

half-finished goods * полуфабрикаты

composites material * композиционные материалы

c) Первый этап формализации запроса

(технологи$3 ADJ производств$3 OR способ$3 ADJ изготовлени$3 OR способ$3 ADJ получени$3 OR способ$3 ADJ производств$3) AND полуфабрикат$3 AND (композиционн$3 ADJ материал$3)

d) Второй этап формализации запроса

(технологи$3 ADJ производств$3 OR способ$3 ADJ изготовлени$3 OR способ$3 ADJ получени$3 OR способ$3 ADJ производств$3) WITH полуфабрикат$3 OR полуфабрикат$3 WITH (композиционн$3 ADJ материал$3)

В табл. 2 приведен еще один пример перевода и формализации запроса, когда переводные эквиваленты слов представлены их различными словообразовательными вариантами.

Таблица 2

Перевод и формализация запроса
Estimation of ocean wave spectra using two-antenna SAR systems

a) Концептуальный анализ запроса

Estimation	00001	оценивание / оценка / расчет / расценка / определение
of	00002	
ocean	00003	00004 океаническая волна/ океанская волна
wave	00004	
spectra	00005	[спектр, спектральный]
using	00006	[использование, использующий, @, @, используя] / [применение, использующий, @, @, используя]
two-antenna	00007	двухантенный
SAR	00008	00009 РЛСА-система
systems	00009	
•	00010	

b) Англо-русский словарь наименований понятий

Estimation * оценивание / оценка / расчет / расценка / определение

ocean wave * океаническая волна / океанская волна

spectra * [спектр, спектральный]

using * [использование, использующий, @, @, используя] / [применение, использующий, @, @, используя]

two-antenna * двухантенный

SAR systems * РЛСА-системой

c) Первый этап формализации запроса

(оценивани$3 OR оценк$3 OR расчет$3 OR расценк$3 OR определени$3) AND (океаническ$3 ADJ волн$3 OR океанск$3 ADJ волн$3) AND (спектр$3 OR спектральн$3) AND (использовани$3 OR использующ$3 OR используя$3 OR применени$3) AND двухантенн$3 AND РЛСА-систем$3

d) Второй этап формализации запроса

((оценивани$3 OR оценк$3 OR расчет$3 OR расценк§3 OR определени$3) WITH (океаническ$3 ADJ волн$3 OR океанск$3 ADJ волн$3) OR (океаническ$3 ADJ волн$3 OR океанск$3 ADJ волн$3) WITH (спектр$3 OR спектральн$3)) SAME (спектр$3 OR спектральн$3) WITH (использовани$3 OR использующ$3 OR используя$3 OR применени$3) OR (использовани$3 OR использующ$3 OR используя$3 OR применени$3) WITH двухантенн$3 OR двухантенн$3 WITH РЛСА-систем$3)

До сих пор мы стремились повысить полноту поиска информации двумя способами:

1) используя квазисинонимы, которые дает система англо-русского перевода в виде перечней переводных эквивалентов английских слов и словосочетаний; 2) используя усечение слов. Наряду с этим можно использовать информационно-поисковые тезаурусы и словари синонимов и гипонимов (более узких по смыслу слов).

Таблица 3

Перевод и формализация запроса
"Disaster caused by bore-like surf beat"

a) Концептуальный анализ запроса

Disaster	00001	бедствие / аварии / катастрофа
caused	00002	00003 вызванных / вызываемый
by	00003	
bore-like	00004	узкий
surf	00005	00006 прибойные биения
beat	00006	
•	00007	

b) Англо-русский словарь наименований понятий

Disaster * бедствие / аварии / катастрофа

caused by * вызванных / вызываемый

bore-like * узкий

surf beat * прибойные биения

c) Первый этап формализации запроса

(бедстви$3 OR авари$3 OR катастроф$3) AND (вызванн$3 OR вызываем$3) AND узк$3 AND прибойн$3 ADJ биени$3

d) Второй этап формализации запроса

((бедстви$3 OR авари$3 OR катастроф$3) WITH (вызванн$3 OR вызываем$3) OR (вызванн$3 OR вызываем$3) WITH узк$3) SAME (узк$3 WITH (прибойн$3 ADJ биени$3))

e) Третий этап формализации запроса

(({бедстви$3 OR бед$3 OR несчаст$3} OR {авари$3 OR крушени$3 OR наезд$3 OR пожар$3 OR пробо$3} OR {катастроф$3 OR авари$3 OR катаклизм$3 OR крушени$3 OR пожар$3}) WITH (вызванн$3 OR вызываем$3) OR (вызванн$3 OR вызываем$3) WITH узк$3) SAME (узк$3 WITH (прибойн$3 ADJ биени$3))

Создание систем поиска информации в политематических базах данных с использованием тезаурусов на сегодняшний день связано с большими трудностями. Во-первых, потому, что тезаурусов мало, а во-вторых, они плохо отражают терминологическое богатство естественных языков и еще хуже парадигматические отношения между терминами. Использование словарей синонимов, гипонимов и гиперонимов более реально. Они меньше по объему, чем тезаурусы, и обеспечивают большее покрытие политематических текстов.

В ВИНИТИ в течение ряда лет создавался политематический словарь синонимов, гипонимов и гиперонимов. В процессе его составления было использовано более 70-ти тезаурусов системы ГАСНТИ и Советский энциклопедический словарь.

Затем был проведен масштабный эксперимент по избыточному индексированию синонимами и гиперонимами рефератов документов из всего спектра тематических областей БД ВИНИТИ. Оказалось, что в результате такого индексирования число возможных входов в тексты рефератов по всем тематикам БД в среднем утроилось. И хотя рассматриваемый словарь далеко не полон ни по лексике, ни по парадигматике, он, на наш взгляд, может служить эффективным средством повышения полноты поиска информации.

В автоматизированных информационно-поисковых системах словарь синонимов, гипонимов и гиперонимов можно применять для двух целей: а) для "избыточного" индексирования поисковых массивов; б) для "избыточного" индексирования запросов. В первом случае потребуется перезагружать базы данных, во втором — этого можно избежать. Авторы статьи пошли по второму пути. В систему автоматического перевода и формализации запросов был включен дополнительный третий этап формализации — этап обогащения слов синонимами и гипонимами. В табл. 3 приведен пример, когда для ряда слов запроса на третьем этапе формализации введены их синонимы и гипонимы (для наглядности цепочки синонимов и гипонимов заключены в фигурные скобки).

Таким образом, описанная система автоматического перевода на русский язык и автоматической формализации англоязычных запросов позволяет проводить поиск информации в русскоязычных базах данных ВИНИТИ, не зная русского языка, и, благодаря наличию на сервере Института системы русско-английского перевода, получать результаты поиска на английском языке.

В заключение следует отметить, что разработка системы поиска информации в русскоязычных базах данных по запросам на английском языке велась при финансовой поддержке Российского Фонда Фундаментальных Исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Белоногов Г. Г., Зеленков Ю. Г., Новоселов А. П., Хорошилов Ал-др А., Хорошилов Ал-сей А. Системы фразеологического машинного перевода. Состояние и перспективы развития. // НТИ. Сер. 2.— 1998.— № 12.

2. Белоногов Г. Г., Зеленков Ю. Г., Новоселов А. П., Хорошилов Ал-др А., Хорошилов Ал-сей А. Метод аналогии в компьютерной лингвистике // НТИ. Сер. 2.— 2000. — № 1.— С. 21.

Материал поступил в редакцию 22.06.2000