

Лингвистика для бизнеса: поиск общего языка



Е. Коржов
директор компании
ТЕКОН

В современной деловой информации текст составляет по разным оценкам от 80 до 90 %. В статье рассмотрены некоторые задачи бизнеса, в решении которых могут помочь лингвистические средства, а также перспективные направления в автоматизации обработки текста.

Когда речь идет о создании перспективных информационных технологий, то на передний план выходят проблемы автоматической обработки текстовой информации, представленной на естественных языках. Это определяется тем, что мышление человека тесно связано с его языком. Более того, естественный язык является инструментом мышления и универсальным средством общения между людьми — средством восприятия, накопления, хранения, обработки и передачи информации.

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА: ПРОБЛЕМЫ И СРЕДСТВА

Решением задач автоматической обработки информации, представленной на естественном языке (письменная и устная речь) занимается компьютерная лингвистика.

С научной точки зрения ее главными проблемами являются:

- ❖ моделирование процесса понимания смысла (то есть перехода от текста к формализованному представлению его смысла, с которым может работать компьютер);
- ❖ синтез речи (обратный переход от формализованного представления смысла к текстам на естественном языке).

Сложность этих проблем вызвана тем, что человеческий язык представляет собой сложную и быстро изменяющуюся систему. В качестве единиц языка и речи могут выступать единицы различного уровня: морфемы, слова, словосочетания, фразы, предложения, сверхфразовые единства (абзацы, параграфы). Эти единицы в совокупности представляют собой иерархическую систему, в которой смысл единиц более высокого уровня не всегда может быть «вычислен» на основе информации о смысле единиц более низкого уровня и информации о связях между этими единицами.

Единицы языка могут связывать различные отношения:

- ❖ синтагматические («и — и») — отношения сочетаемости элементов одного уровня;
- ❖ парадигматические («или — или») — отношения противопоставленности элементов одного уровня;
- ❖ иерархические — отношения вхождения более простой единицы языка в более сложную (например, слова в предложение).

Средства, создаваемые и применяемые в компьютерной лингвистике, можно условно разделить на *декларативные* (словари единиц языка и речи, тексты и различного рода грамматические таблицы), и *процедурные* (средства манипулирования этими элементами).

Важной проблемой прикладной компьютерной лингвистики является оценка необходимого соотношения между декларативной и процедурной компонентами систем автоматической обработки текстовой информации.

Чему отдать предпочтение: мощным вычислительным процедурам, опирающимся на относительно небольшие словарные системы с богатой грамматической и семантической информацией, или мощной декларативной компоненте при относительно простых процедурных средствах? Многие считают второй путь предпочтительным, что подтверждается полувековым опытом развития компьютерной лингвистики. То есть успех в решении прикладных задач компьютерной лингвистики зависит, прежде всего, от полноты и точности представления в компьютере декларативных средств.

Далее рассмотрим основные задачи лингвистического обеспечения процессов сбора, накопления, обработки и поиска текстовой информации.

ОБНАРУЖЕНИЕ И ИСПРАВЛЕНИЕ ОШИБОК ТЕКСТОВ

Эта задача может быть условно разделена на три подзадачи — орфографического, синтаксического и семантического контроля текстов.

Орфографический контроль обеспечивается с помощью процедуры морфологического анализа, использующей эталонный машинный словарь основ слов. В процессе орфографического контроля слова текста подвергаются морфологическому анализу, и если их основы отождествляются с основами эталонного словаря, то они считаются правильными; если не отождествляются, то они в сопровождении «окружения» выдаются на просмотр человеку (процедура, достаточно знакомая всем по работе с MS Word).

Синтаксический контроль текстов существенно сложнее. Во-первых, он включает в себя и орфографический контроль, во-вторых, проблема синтаксического анализа неформализованных текстов еще не решена в полном объеме. Здесь можно идти двумя путями: (1) составлять машинные словари эталонных синтаксических структур и сравнивать с ними структуры анализируемого текста; (2) разрабатывать сложную систему правил проверки грамматической согласованности элементов текста.

Семантический контроль с целью обнаружения смысловых ошибок относится к классу задач искусственного интеллекта и в полном объеме может

быть обеспечен только на основе моделирования процессов человеческого мышления. Для этого придется создавать мощные энциклопедические базы знаний и программные средства манипулирования знаниями. Тем не менее, для ограниченных предметных областей и для формализованной информации эта задача вполне разрешима.

Как видим, для решения этой (и многих других) задач необходимы словари. Поэтому представляется перспективным такой путь развития компьютерной лингвистики, когда основные усилия будут направлены на создание мощных словарей, изучение их семантико-синтаксической структуры и на создание базовых процедур анализа и синтеза текстов. Такие словари станут фундаментом для решения широкого спектра прикладных задач.

Подобный словарь используется в продукте **Ukrainian Context Optimizer (UCO)**, который был создан фирмой ТЕКОН совместно с ее партнерами «ЭР СИ О» (Россия) и «Трайидент Софтвр» (Украина). Его общий объем — более 115 тысяч слов (около 4 миллионов словоформ). Благодаря эффективной системе описания украинской морфологии, объем хранимых лингвистических данных и данных, необходимых для анализа неизвестных слов, минимален (менее 10 Мб).

СОСТАВЛЕНИЕ СЛОВАРЕЙ

Машинные словари являются неотъемлемой частью любой системы автоматической обработки текстовой информации. Они могут представлять собой словари слов и словосочетаний, выражающих характерные понятия для той или иной области деятельности.

Полнота отображения. При составлении словарей необходимо стремиться к тому, чтобы они в максимальной степени отражали лексический (словарный) состав текстов. Поэтому для их составления необходимы тексты достаточно большого объема (как минимум, в несколько десятков миллионов лексических единиц). А такая работа может быть выполнена в разумные сроки только при использовании средств автоматизации.

Задача составления словарей словосочетаний, выражающих понятия, гораздо сложнее задачи составления словарей слов, поскольку словосочетания в тексте формально не выделены, а их границы «отмечены» лишь в сознании человека (то есть, элементы словосочетания могут менять порядок, быть опущены или содержать «промежуточные» слова и т.п.).

Установка отношений и связей. В системах автоматической обработки текстов важно установить отношения между понятиями, выраженными отдельными словами или словосочетаниями (вроде

«одинаковость (синонимия)», «род-вид», «целое-часть», «причина-следствие» и других). Это крайне трудоемкая задача.

Автоматизация выявления отношений между отдельными словами была бы сравнительно легким делом, если бы мы располагали формализованными описаниями «смыслов» слов, где каждое слово характеризовалось бы набором семантических признаков. Но такие описания практически отсутствуют. Многочисленные попытки их составления носили, как правило, экспериментальный характер и завершались составлением не особо представительных семантических словарей, объемом порядка нескольких тысяч лексических единиц и недостаточно глубоким описанием.

Но если мы не располагаем достаточно представительными формализованными описаниями слов, то их неформализованные или слабо формализованные описания представлены в различного рода **словарях**: толковых (объясняющих), терминологических (уточняющих) и энциклопедических (объединяющих). В этих словарях определяемое слово (или словосочетание) обычно соотносится с лексической единицей, выражающей более широкое (родовое) понятие и с лексическими единицами, характеризующими ее отличительные признаки.

Другим источником выявления отношений между словами могут служить **тезаурусы** (от греч. сокровище) — разновидность словарей общей или специальной лексики, содержащих слова вместе с их смысловым «окружением». Слова классифицированы по тематическим категориям (рубрикам) и синонимическим рядам (для каждого слова установлены его синонимы, более общие и более частные понятия, а также «родственные» слова, часто имеющие с ним смысловую связь в тексте). В отличие от толкового словаря, тезаурус позволяет выявить смысл слова не только с помощью его определения, но и через соотнесение слова с другими понятиями и их группами. Тезаурусы, особенно в электронном формате, являются одним из действенных инструментов для описания отдельных предметных областей.

ИНДЕКСИРОВАНИЕ ДОКУМЕНТОВ И ЗАПРОСОВ

Решение этой задачи необходимо для ускорения поиска нужной информации (иначе пришлось бы просматривать все тексты всех документов).

Поначалу под индексированием понимали процесс присвоения документам и запросам классификационных индексов, отражающих содержание документов. В дальнейшем это понятие трансформировалось, и термином «индексирование» стали называть процесс перевода описа-

ний документов и запросов с естественного языка на формализованный, в частности, на язык «поисковых образов».

Поисковые образы документов (ПОД), как правило, оформляются в виде перечней ключевых слов и словосочетаний, отражающих их тематическое содержание. Процесс индексирования документов связан со значительными затратами квалифицированного труда. Поэтому уже в 60-х годах прошлого столетия стали предприниматься попытки автоматизировать этот процесс.

Построение индекса для всего информационного пространства организации — задача крайне сложная и объемная. Поэтому для ее упрощения в некоторых случаях применяют *классификацию* (разбиение множества документов по отдельным группам и подгруппам — как компьютерных файлов по каталогам и подкаталогам) и *реферирование* (создание краткого описания, достаточно точно отражающего информацию «основного» документа)

Реферирование документов. Самой трудной задачей, которую приходится решать и при ручном, и при автоматическом индексировании, является правильный выбор перечней понятий, которыми следует описывать смысловое содержание документов («системы координат»). Для такой работы требуется высокая квалификация ее исполнителей.

Автоматическое индексирование документов удобно проводить по текстам их рефератов, поскольку обычно они составляются квалифицированными специалистами, умеющими изложить основное содержание документов в концентрированном виде. Тогда процесс индексирования документов можно свести к автоматическому анализу текстов их рефератов и составлению ПОДов путем формирования списка различных наименований понятий, выявленных в результате такого анализа.

Автоматизация реферирования документов важна для крупных организаций, которым решение этой задачи позволило бы сберечь значительные средства. Ведь рефераты документов, составляемые квалифицированными специалистами, содержат не только краткий пересказ их содержания, но, как правило, и их оценку. Автоматизировать процесс оценки содержания документов пока еще очень трудно.

Эксперименты показали, что ПОДы, составленные в автоматическом режиме по заголовкам и рефератам документов, обеспечивают большую полноту поиска, чем ПОДы, составленные вручную. Объясняется это тем, что система автоматического индексирования более полно отражает различные аспекты содержания документов, чем система ручного индексирования.

Поисковые образы запросов оформляются в виде логических конструкций, в которых ключевые слова и словосочетания соединялись друг с другом логическими и синтаксическими операторами.

При автоматическом индексировании запросов возникают примерно те же проблемы, что и при автоматическом индексировании документов. Здесь также приходится выделять ключевые слова и словосочетания из текста и нормализовать слова, входящие в текст запроса. Логические связи между ключевыми словами и словосочетаниями могут проставляться вручную или с помощью автоматизированной процедуры. Важным элементом процесса автоматического индексирования запроса является дополнение входящих в его состав ключевых слов и словосочетаний.

Например, в Oracle Text, из средств расширения поискового запроса можно выделить три группы. Во-первых, это расширение слов запроса всеми морфологическими формами, что реализуется привлечением знаний о морфологии языка. Во-вторых, допускается расширение слов запроса близкими по смыслу словами за счет подключения тезауруса. И последнее — это расширение запроса словами, близкими по написанию и по звучанию (нечеткий поиск и поиск созвучных слов). Нечеткий поиск целесообразно применять при поиске слов с опечатками, а также в тех случаях, когда возникают сомнения в правильности написания фамилии, названия организации и т.п. Поскольку в Oracle Text все эти возможности не распространяются на восточноевропейские языки, для обработки документов на украинском языке необходимо применять Ukrainian Context Optimizer

ПРОБЛЕМЫ ПОИСКА ИНФОРМАЦИИ

Проблемы поиска информации уже частично рассматривались выше в связи с задачей автоматического индексирования. Однако более перспективным является поиск документов по их полным текстам, так как использование для этой цели «заменителей» (библиографических описаний, поисковых образов документов, рефератов) обычно приводит к потере информации при поиске. Как показывает практика, наибольшие потери происходят тогда, когда вместо первичных документов используются их библиографические описания, наименьшие — при использовании рефератов.

Характеристики поиска. С появлением автоматизированных документальных поисковых систем возник вопрос о качестве поиска. Дело в том, что часть документов, выдаваемых потребителю в результате автоматического поиска, оказывалась нерелевантной запросу (не отвечающей ему «по

смыслу»), а часть релевантных документов, содержащихся в поисковом массиве, ему не выдавалась. Первое явление получило название «поисковый шум», второе — «потери информации». Для количественной оценки этих явлений были введены понятия *коэффициент шума* (отношение количества нерелевантных документов к общему количеству документов, выданных в результате поиска), и *коэффициент потерь* (отношение количества релевантных документов, не найденных в поисковом массиве, к общему количеству таких документов, содержащихся в поисковом массиве).

Важными характеристиками качества поиска информации являются его **полнота и точность**. Полнота поиска может быть обеспечена путем максимального учета парадигматических связей между словами и словосочетаниями, а точность — путем учета их синтагматических связей. Были введены также понятия коэффициента точности поиска и коэффициента его полноты. Значение *коэффициента точности* полагалось равным дополнению к единице значения коэффициента шума, а значение *коэффициента полноты* — дополнению к единице значения коэффициента потерь.

Существует мнение, что полнота и точность поиска находятся в обратной зависимости: меры по улучшению одной из этих характеристик приводят к ухудшению другой. Но это справедливо только для фиксированной логики поиска. Если эту логику удастся совершенствовать в ходе поиска, то обе характеристики могут улучшаться одновременно.

Подходы к решению. Проблему обеспечения полноты и точности поиска информации разработчики автоматизированных поисковых систем пытаются решать различными методами.

Одним из них является **метод ранжирования** выдаваемых документов. По этому методу на основе исходного поискового запроса генерируется ряд других запросов с ослабленными условиями поиска, а найденные документы упорядочиваются по убыванию предполагаемой степени их релевантности исходному запросу. При этом у пользователя имеется возможность просматривать не все найденные документы, а только ограниченное их число.

Другой метод решения проблемы обеспечения полноты и точности поиска состоит в использовании **концепции гипертекста**. Обычно гипертекст определяется как технология работы с текстовыми данными, позволяющая устанавливать «гиперсвязи» между отдельными терминами, фрагментами документов и статьями в текстовых массивах. Благодаря этому обеспечивается не только последовательная («линейная») работа с текстом, как при обычном чтении, но и произвольный доступ к информации и ее просмотр в соответствии с установленной структурой связей.

Гипертекстовые связи представляют собой по существу перекрестные ссылки, позволяющие мгновенно обращаться к нужным фрагментам информации. Они наиболее эффективны тогда, когда используются при поиске в больших массивах информации, разделенных на множество мелких связанных по смыслу фрагментов, и когда пользователю в каждый данный момент требуются только небольшие объемы информации.

Инверсная форма. На рубеже 70-х и 80-х годов прошлого столетия появились программные системы, в которых текстовые файлы представлялись в компьютере одновременно в прямой и в инверсной форме. *Прямая форма* представления текстов — это обычная их запись, подобная той, которая используется на бумажных носителях информации. В *инверсной форме* текст представляется в виде алфавитного списка всех входящих в него слов (кроме служебных частей речи — местоимений, предлогов, союзов и пр.) с указанием для каждого слова адресов его «вхождения» в текст. Инверсная форма занимает значительный объем памяти, но существенно расширяет поисковые возможности автоматизированных систем.

Инверсные файлы и гипертекстовое представление информации часто используются совместно, в одной и той же поисковой системе. При этом инверсные файлы обеспечивают начальное обращение к фрагментам текстов по запросам, а гипертекст дает возможность продолжать поиск, используя ассоциативные связи между этими фрагментами.

Диалог. Процесс поиска информации в полнотекстовых базах данных целесообразно строить как процесс диалогового общения пользователя с информационно-поисковой системой, при котором человек последовательно просматривает фрагменты текстов (абзацы, параграфы), отвечающие логическим условиям запроса, и отбирает те из них, которые представляют для него интерес. В качестве окончательных результатов поиска могут выдаваться как полные тексты документов, так и любые их фрагменты.

Примером такой организации поиска является **UOSES** (Ukrainian Optimized Secure Enterprise Search) — многоплатформенный программно-аппаратный комплекс, предназначенный для защищенного поиска информации в условиях корпоративного использования, в частности решения задач бизнес-анализа и бизнес-разведки. Он разработан компанией ТЕКОН в сотрудничестве с «ТрайидентСофтвр» и является развитием ROSES (Russian Optimized Secure Enterprise Search) — решения, созданного российскими партнерами ТЕКОН, компаниями «ФОРС — Центр разработки» и «ЭР СИ О».

ТЕМАТИЧЕСКИЙ ПОИСК

Неоценимую помощь при поиске может оказать классификация документов по темам (например, в случае, если пользователь затрудняется точно подобрать ключевые слова, или же хочет сузить область поиска, уточнив тематику, по которой следует искать документы). Это обеспечивает тематическому поиску более высокую точность и полноту по сравнению с обычным контекстным поиском. Кроме того, он позволяет найти документы, вовсе не содержащие слов из названия заданной темы, однако имеющие к ней отношение.

Для автоматизированного отнесения документа к тем или иным темам можно воспользоваться тезаурусом (как, например, реализуется в Oracle Text).

Следует отметить уникальную способность Ukrainian Context Optimizer устанавливать смысловые связи между темами, выявляя их в тексте. В отличие от предопределенных и достаточно очевидных связей, которые задаются в тезаурусе, УСО выявляет динамические связи, большинство из которых уникальны для каждой коллекции документов. Это позволяет найти в коллекции документов совокупность тем, связанных по смыслу со словами запроса.

ЛИНГВИСТИЧЕСКИЕ ПРОЦЕССОРЫ

Перспективной задачей компьютерной лингвистики является построение лингвистических процессоров, обеспечивающих общение пользователей с автоматизированными информационными (в частности с экспертными системами) на естественном языке или на близком к естественному. Лингвистические процессоры исполняют роль посредников между человеком и компьютером и должны решать следующие основные задачи:

- ❖ переход от текстов входных информационных запросов и сообщений на естественном языке к представлению их смысла на формализованном языке;
- ❖ переход от формализованного представления смысла выходных сообщений к его представлению на естественном языке.

Первая задача должна решаться путем морфологического, синтаксического и концептуального анализа входных запросов и сообщений, вторая — в обратном порядке — путем концептуального, синтаксического и морфологического синтеза выходных сообщений.

Концептуальный анализ информационных запросов и сообщений состоит в выявлении их понятийной структуры (границ наименований понятий и отношений между понятиями в тексте) и переводе этой структуры на формализованный язык. Он проводится после морфологического и синтаксического

анализа запросов и сообщений. Концептуальный синтез сообщений состоит в переходе от представления элементов их структуры на формализованном языке к вербальному (словесному) представлению. После этого сообщениям придается необходимое синтаксическое и морфологическое оформление.

Для функционирования лингвистических процессоров необходимо иметь в их составе процедуры морфологического, синтаксического и концептуального анализа и синтеза текстов, а также базу знаний, содержащую словари единиц языка и речи и их синтагматические и парадигматические характеристики. Эффективность лингвистических процессоров зависит не только от качества процедурных средств, но и от качества лингвистической базы знаний: насколько адекватно и полно представлено в ней многообразие явлений естественного языка. А качественную лингвистическую базу знаний можно создать только на основе широкого применения средств автоматизации.

ИЗВЛЕЧЕНИЕ ФАКТОГРАФИЧЕСКОЙ ИНФОРМАЦИИ

Базы знаний организаций трудно создавать без использования средств автоматизации. Важную роль будут играть лингвистические процессоры, особенно при автоматизированном извлечении из неформализованных текстов фактографической информации (данных в виде объектов и связей).

Для автоматизации подобных задач, компания ТЕКОН предлагает **Fact Extractor Ukraine** — интеллектуальную программу для высокоточного избирательного анализа текстов на украинском языке и выявления фактов различного типа, связанных с заданными объектами (персоны, организации, география, предметы, действия, атрибуты и др.). Основная сфера применения — задачи из области компьютерной разведки, требующие точного поиска информации (например, автоматический подбор материала к досье на целевой объект или мониторинг определенных сторон его активности, освещаемых в СМИ). Например, можно не только найти фрагменты текста, в которых говорилось о поездках персоны, ее встречах, заключении договоров, сделках купли-продажи, но и точно определить все места поездок, визави и контрагентов, наименование товаров и прочее.

Программа работает в среде Windows (2000 и выше) и позволяет обрабатывать документы в основных текстовых форматах из различных источников — файловые системы, web-сайты, базы данных. Результат работы — таблица, которая содержит информацию о найденных фактах и может экспортироваться в html-формат для формирования отчета или для загрузки в стороннее приложение.

Fact Extractor Ukraine предполагает настройку шаблонов для поиска и классификации фактов.

Стандартные шаблоны, включенные в комплект поставки, позволяют распознавать огромное количество разнообразных фактов, но без детальной классификации (т. е. просто находить события, в которых участвует указанный объект, и извлекать из текста фигурантов этих событий без детализации их ролей). Специализированные шаблоны либо приобретаются отдельно, либо создаются пользователем при помощи дополнительной программы **Fact Tuner**.

Есть также инструмент для разработчиков (**Fact Extractor SDK**), на базе которого построен **Fact Extractor** и который позволяет включать возможности анализа текста в собственные приложения. В него, помимо общих словарей и правил языка, входят правила выделения специальных объектов (дат, адресов, документов, телефонов, денежных сумм, марок автомобилей и пр.), шаблоны для распознавания различных классов событий и фактов (сделок, экономических показателей, конфликтов, биографических фактов и пр.), характеристик объекта (позитива, негатива и др.), высказываний прямой и косвенной речи. Также в состав SDK входят исходные коды приложений на C++, иллюстрирующие использование библиотеки для решения ряда типовых задач, например: построения смыслового портрета документа (множества слов и словосочетаний, ранжированных по значимости); построения реферата текста (в том числе рефератов по каждой сущности); построения иерархического глоссария по коллекции текстов; трансляции запроса на естественном языке в пакет запросов поисковой машины.

Fact Extractor SDK работает на платформах Windows и Unix.

МАШИННЫЙ ПЕРЕВОД ТЕКСТОВ

Как видно из предыдущих рассуждений, при автоматическом поиске информации приходится преодолевать языковой барьер, возникающий между пользователем и поисковой системой в связи с имеющим место в текстах разнообразием форм представления одного и того же смысла. Этот барьер становится еще более значительным, когда поиск приходится вести в многоязычных базах данных. Кардинальным решением проблемы здесь мог бы стать машинный перевод текстов документов с одних языков на другие. Это можно делать либо заранее, перед загрузкой документов в поисковую систему, либо в процессе поиска информации. В последнем случае запрос пользователя должен переводиться на язык массива документов, в котором ведется поиск, а результаты поиска — на язык запроса. Такого рода поисковые системы уже доступны в Internet.

Евгений Коржов