Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus

© Тихомиров И. А.

Институт Системного Анализа РАН
matandra@isa.ru

Аннотация

В статье описаны отличительные особенности вопросно-ответного поиска. Приводится краткое описание ситуативно-реляционной модели и ее применения для решения задачи вопросно-ответного поиска. Показаны базовые принципы работы интеллектуальной поисковой системы Exactus и сделаны выводы о направлениях ее дальнейшего развития.

1. Отличительные особенности вопросоно-ответного поиска

Вопросно-ответный поиск имеет существенные отличия от поиска по ключевым словам или его аналогов. Вопрос к поисковой системе формируется, как правило, в виде вопросительного предложения. Оно содержит как известную информацию об объекте, его свойстве или сопровождающем его явлении (список далеко не полон), так и «подразумеваемый» фрагмент отсутствующей информации, выраженный, в частности, в виде вопросительного слова. Например, пользователь знает имя некоторого человека, но не знает дату его рождения. В результаты будет сформирован следующий запрос к поисковой системе: «когда родился Есенин?».

Основная отличительная особенность вопросно-ответного поиска заключается в поиске слова, словосочетания или целого предложения, которое не содержится в явном виде в запросе. Вместо этого имеется информация «намекающая» поисковой системе о возможном ответе на вопрос пользователя.

Вопросно-ответный поиск имеет другую цель по сравнению с обычным поиском. Его задача – найти фрагмент документа, содержащий точный и достоверный ответ на поставленный вопрос.

Очевидно, что при реализации функций вопросно-ответного поиска приходится иметь дело с естественным языком, а именно законченными фразами или предложениями, сформированными по определенным законам. Соответственно, требуется применение адекватных лингвистических средств по работе с естественным языком - анализаторов различного вида. Выбор анализаторов зависит от модели, выбранной разработчиками поисковой системы для реализации функций вопросно-ответного поиска.

В данной статье приводится описание ситуативно-реляционной модели текста, описывающей его семантику.

2. Ситуативно-реляционная модель текста

Обратимся к теории коммуникативной грамматики русского языка, которая представляет собой оригинальный подход к описанию русского синтаксиса [1,2]. Эта теория опровергает традиционное противопоставление синтаксиса семантике, которое предполагает разделение знаний о законах формирования связной речи на два уровня: знания о форме (синтаксис) и знания о значении (семантика). Основополагающая идея коммуникативной грамматики заключается в том, что синтаксис должен изучать именно осмысленную речь, а синтаксические правила должны учитывать категориальные значения слов, чтобы иметь возможность определять обобщенный смысл любой синтаксической конструкции – от слова до словосочетания и простого предложения.

Очевидно, что одних морфологических характеристик недостаточно, чтобы слово стало конструктивной единицей синтаксиса. Слово-лексема еще не является синтаксической единицей, слово – единица лексики, а в разных его формах могут реализоваться или актуализироваться разные стороны его общего значения. Формируя и изучая связную речь, синтаксис имеет дело прежде всего с осмысленными единицами, несущими свой не индивидуально-лексический, а обобщенный, категориальный смысл в конструкциях разной степени сложности. Обобщенное, или категориальное, значение определяет синтаксические возможности слова и способы его функционирования.

Определение 1:
Синтаксемой называется минимальная синтактико-семантическая единица языка, несущая свой обобщенный категориальный смысл в конструкциях разной степени сложности и характеризующаяся всегда взаимодействием морфологических, семантических и функциональных признаков.

Несмотря на сложность описания, *синтаксема* является интуитивно понятной конструкцией для любого носителя языка и используется им повсеместно для построения различного рода высказываний. Пример: «Митрофанушка не знал, что говорит прозой». Синтаксемами имен существительных в данном предложении являются: «Митрофанушка» – личное существительное именительного падежа, принадлежащее к классу имен собственных и играющее в данном предложении роль субъекта. «Прозой» - существительное в творительном падеже, принадлежащее к классу признаковые и играющее в данном предложении роль медиатива.

Для понимания текста одних синтаксем недостаточно. Важна именно их сочетаемость друг с другом в конкретном предложении. Эта сочетаемость определяется множеством отношений [3].

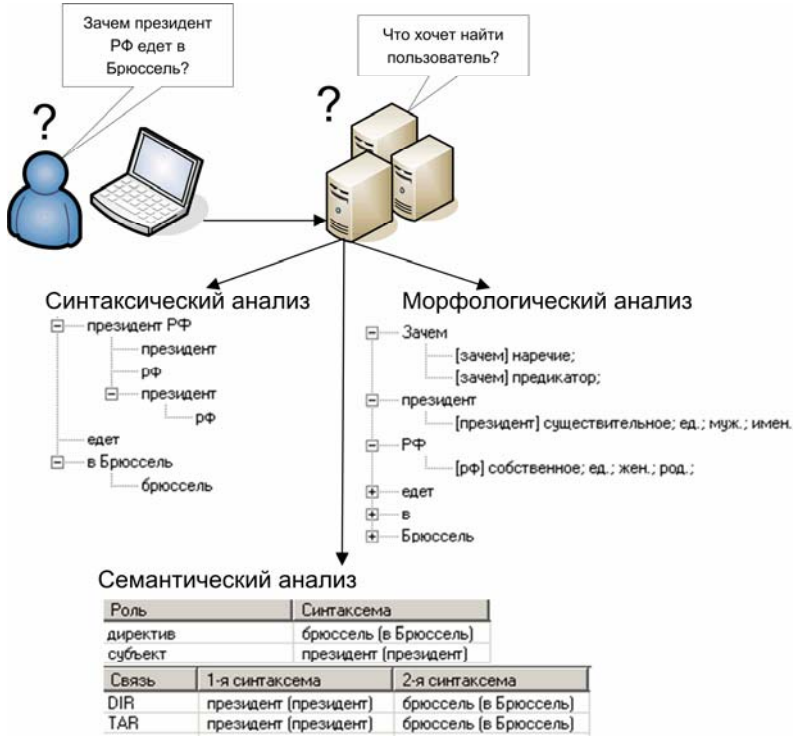
Определение 2:
Смысл предложения определяется совокупностью входящих в него синтаксем и множеством отношений на них.

Для определения синтаксем и множества отношений на них требуется анализ текста различного вида, а именно: морфологический, синтаксический и семантический анализ [4].

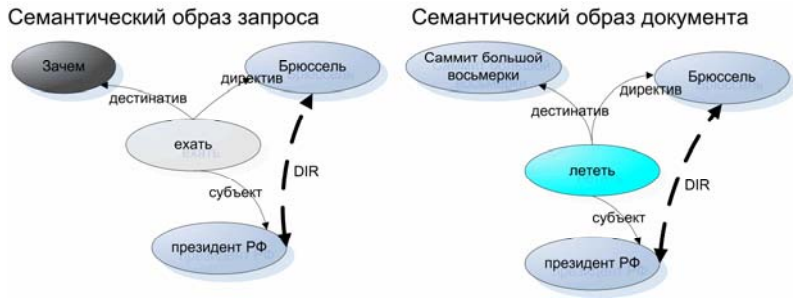
3. Применение ситуативно-реляционной модели текста для реализации функций вопросно-ответного поиска

Рассмотрим применение ситуативно-реляционной модели текста на примере: «Зачем президент РФ едет в Брюссель?».

Как уже отмечалось выше, для построения модели предложения требуется провести морфологический, синтаксический (определение главного слова в именной группе), семантический анализ текста [4]. В случае вопросно-ответного поиска это анализ запроса и анализ документов коллекции текстов.



В результате анализа запросов и текстов документов строится семантические сети – образы запроса и документов. Для каждого предложения документа коллекции строится свой образ:



Для реализации функции поиска необходимо определить операцию сравнения образов – алгоритм вычисления релевантности. Exactus находит документы, семантические образы которых наибо-

лее близки семантическому образу запроса по синтаксемам, связям между ними и конкретным значениям вершин сети (вне зависимости от формы выражения). Приоритет имеет именно сравнение синтаксем и связей, а не совпадение конкретных словоформ запроса и найденных документов.

Следует отметить, что режим вопросно-ответного поиска реализуется в Exactus «естественным образом», то есть является штатным режимом поиска, так же, как и другие виды поиска: ситуационный, фактографический и т.д. Это означает, что при всех видах поиска применяются те же самые алгоритмы, но с разными настроечными параметрами [4].

4. Заключение

В результате участия в семинаре РОМИП'2006 удалось проверить работоспособность поисковых алгоритмов Exactus при вопросно-ответном поиске с использованием независимых экспертов. Предварительный анализ результатов показал направления дальнейшего развития Exactus, а именно:

1. Проверка достоверности результатов ответов на вопросы по статистическим критериям. Для Exactus «Америку открыли китайцы» и «Америку открыл Колумб» - одинаково релевантные ответы. С точки зрения здравого смысла первый ответ является недостоверным, хотя в коллекции он присутствовал.

2. Отсечение слаборелевантных результатов при поиске. Exactus всегда выдавал по 10 первых результатов в списке, даже если документы были слаборелевантны.

Кроме того, разработчики Exactus хотели бы предложить возможные варианты совершенствования методики оценки вопросно-ответной дорожки [5]. Среди модификаций методики разработчики предлагают не учитывать запросы, предоставленные участником, при проверке и сопоставлении результатов. Подобный подход может повысить достоверность результатов.

В заключение разработчики Exactus хотели бы выразить признательность организационному комитету РОМИП за организацию семинара, позволяющего разработчикам поисковых систем в России и за ее пределами оценить качество создаваемых систем в различных аспектах их работы.

Литература

[1] Золотова Г.А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. Институт русского языка РАН им. В. В. Виноградова, М. 2004 – 544 с.

[2] Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. – М.: Наука, 1988 – 440 с.

[3] Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997.

[4] Осипов Г.С., Завьялова О.С., Климовский А.А., Кузнецов И.А., Смирнов И.В., Тихомиров И.А. Проблемы обеспечения точности и полноты поиска: Пути решения в интеллектуальной метапоисковой системе "Сириус". //Труды мезденародной конференции Диалог'2005, с. 390-395, Москва, Наука, 2005.

[5] Осипов Г.С., Выборнова О. Е., Завьялова О.С., Смирнов И.В., Тихомиров И.А. Методика оценки эффективности систем информационного поиска// Сборник трудов VI международной конференции Интеллектуальный Анализ Информации ИАИ‘2006, г. Киев, стр. 214-227.

QA search in intelligent search engine Exactus

Ilya A. Tikhomirov
matandra@isa.ru

The paper presents specialties of QA search. Paper describes application of situate-relevant search model to QA search. We state the effectiveness of Exactus system based on semantic search involving refined linguistic processing tools. Future work description and conclusion are given.