Н.Ю. Шкуратова (Минск, МГЛУ)

ПРИНЦИПЫ ОБРАБОТКИ ЗАПРОСОВ ПОЛЬЗОВАТЕЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ ПОИСКОВОЙ СИСТЕМОЙ

Поиск информации – задача, которую человечество решает уже многие столетия. По мере роста объема информационных ресурсов, потенциально доступных человеку, были выработаны все более изощренные и совершенные поисковые средства и приемы, позволяющие найти необходимый документ. Основная задача любой поисковой системы – организовать доступ пользователя к любому информационному ресурсу. Известно, что поисковые системы (как и алгоритмы поиска) появились задолго до распространения глобальной сети Интернет. Но именно популярность Сети и тот факт, что ею стали постоянно пользоваться люди, не имеющие специального образования и вообще слабо разбирающиеся в компьютерах, стали толчком для активного развития поисковых систем. И если двадцать лет назад рассуждения об интерпретации компьютером запросов, составленных на естественном языке, были теоретическими абстракциями, то сегодня проблема интерпретации запроса является актуальной. Переучить пользователя, научить его правильно с точки зрения поисковой системы составлять запросы практически невозможно. Приходится идти с другой стороны – пытаться научить систему понимать то, что хочет найти человек. «Интеллектуальность» современных поисковых систем позволяет во многих случаях определить тему и цель поиска благодаря различным технологиям лингвистического анализа запроса. Есть предпосылки полагать, что в скором будущем будет создана универсальная интеллектуальная поисковая система, алгоритм работы которой сможет удовлетворить в поиске информации любого конечного пользователя.

Принимая во внимание, что на сегодняшний момент чаще всего информация представляется в текстовом виде, будем рассматривать описанные технологии и стратегии поиска применительно к этому типу данных. Поиск таких видов ресурсов как мультимедиа или графическая информация не столь сложен, не столь многогранен и во многих случаях опирается на технологии, используемые при поиске текстовой информации.

Материалом для исследования послужили тексты 200 англоязычных запросов в виде вопросительных предложений на тему «Современные информационные технологии», взятые из баз данных поисковых систем глобальной сети Интернет. Их предварительный анализ позволил выделить следующие четыре типа синтаксических структур вопросительных высказываний:

1. What + is (are) + подлежащее + (обстоятельство места / времени)?

Например: *What is a computer virus? What are the future trends in mailware?*

2. What + подлежащее + сказуемое + (дополнение) + (обстоятельство места / времени)?

Например: *What kinds of keyboards are there? What kind of files can spread computer viruses?*

3. How + сказуемое (в форме инфинитива) + дополнение + (обстоятельство места / времени)?

Например: *How to delete an Email from the outbox automatically in Outlook 2007? How to add administrative tools to start menu in Windows 7?*

4. What (How / Where / Why) + вспомогательный (модальный) глагол + подлежащее + сказуемое + дополнение + (обстоятельство места / времени)?

Например: *How can I make my Compaq notebook run faster? How do I speed up Vista?*

Автоматическая сегментация вопросительных предложений связана с выделением в них формальных групп подлежащего, сказуемого, дополнения и обстоятельства места или времени. Эта процедура происходит с опорой на порядок слов, характерный для предложений данного типа, а также на левые и правые граничные сигналы каждой синтаксической группы. В результате лингвис-тического анализа левой и правой контактной дистрибуции членов английского вопросительного предложения были получены списки маркеров левой и правой границ каждого синтаксического сегмента. К ним относятся:

1. Единичные словоформы, общие для левой и правой границы (смешанные границы).

2. Бинарные сочетания, общие для левой и правой границы (смешанные границы).

3. Единичные словоформы, характерные только для левой (или для правой) границы.

4. Бинарные сочетания, характерные только для левой (или для правой) границы.

Например, граничные сигналы, позволяющие выделять группу дополнения в английском предложении, выглядят следующим образом:

Таблица 1

Граничные сигналы для выделения группы дополнения

Левая граница		**Правая граница**	
Эксклюзивная	**Инклюзивная**	**Инклюзивная**	**Эксклюзивная**
Правая граница группы предиката или объекта	Предлоги (*of*)	Существительное	Левая граница следующей группы
Сочинительные и подчинительные союзы	Определенный и неопределенный артикли	———	Сочинительные и подчинительные союзы
Глаголы в личной форме	Местоимения (кроме личных, *who* и *which*)	———	Определенный и неопределенный артикли
Знаки препинания (запятая в случае	Существительное в функции		Предлоги (простые и

однородных объектов, двоеточие в случае перечисления)	определения к следующему за ним существительному	———	сложные)
———	Причастия.	———	Причастие II
———	Словоформы с окончанием *–ing* (герундий и причастие I)	———	Местоимения (включая личные, *who* и *which*)
———	Прилагательные	———	Знаки препинания (все)
———	Числительные	———	Словоформа *to*
	Наречие при прилагательном или причастие II, которые входят в именную группу и определяют существительное		Наречия

Целью лингвистической обработки запроса пользователя на естественном языке является его преобразование интеллектуальной поисковой системой в некоторый формальный семантический вид, например, фрейм. Фрейм – это расположение данных в виде иерархической структуры. Начало этой структуры представлено в виде фрейма высшего уровня (фрейма класса или протофрейма), содержащего наиболее общий вид какой-либо информации. Затем следуют так называемые «дочерние» фреймы, представляющие собой наиболее крупные разновидности этой информации. В конце структуры содержатся фреймы-экземпляры. Они содержат детальное описание каких-то конкретных элементов информации.

В зависимости от содержащейся во фреймах информации они делятся на фреймы-описания, например:

ОВОЩИ [<ПОМИДОРЫ>; <ПЕРЕЦ>; <БАКЛАЖАНЫ>; …]

и ролевые фреймы, например:

ПОХИЩАТЬ [<КТО>; <КОГО>; <С ПОМОЩЬЮ ЧЕГО>; <ОТКУДА>; <КУДА>; <ЗАЧЕМ>; …].

В данной работе в качестве формального способа представления содержания запроса пользователя используются ролевые фреймы. Имя фрейма будет представлено извлеченной на этапе автоматической синтаксической сегментации группой сказуемого в текстовой форме, а в качестве слотов будут выступать группы подлежащего, дополнения и обстоятельства:

СКАЗУЕМОЕ [<группа подлежащего>; <группа дополнения>; <группа обстоятельства>].

Рассмотрим в качестве примера запрос пользователя *How to add administrative tools to start menu in Windows 7?* В ходе автоматического синтаксического анализа этого предложения компьютер выделит следующие члены предложения: *to add, administrative tools, to start menu, in Windows 7*. Далее он распределит полученные синтаксические сегменты в слотах фрейма:

TO ADD [<administrative tools>; <to start menu>;<in Windows 7>].

Сформированная фреймовая структура, передающая основное содержание запроса пользователя, используется поисковым инструментом для поиска релевантных данных в информационной системе.