М. А. Шумская
Минск, МГЛУ

ФУНКЦИОНИРОВАНИЕ И ПУТИ СОВЕРШЕНСТВОВАНИЯ АНГЛОЯЗЫЧНОЙ ВОПРОСНО-ОТВЕТНОЙ ПОИСКОВОЙ СИСТЕМЫ

Рассматривается процедура функционирования экспериментальной англоязычной интеллектуальной поисковой системы, а также предлагаются возможные пути ее совершенствования. Разработанная формальная модель состоит из двух частей. Первая часть – анализатор запроса отвечает за анализ запроса пользователя, представленного в виде специального вопроса, начинающегося с вопросительного слова *When* или *Where*. Вторая часть – синтезатор ответа, опираясь на результаты поиска, синтезирует конкретный ответ на поставленный вопрос. В основу работы полной формальной модели интеллектуальной поисковой системы положена достаточно обширная лингвистическая база данных.

В настоящее время проблема организации релевантного поиска информации в сети Интернет еще до конца не решена, поэтому большое значение имеет корректное распознавание смысла запроса пользователя на естественном языке. На данный момент многие поисковые системы недостаточно хорошо работают с естественно-языковыми запросами и, как правило, используют набор специальных средств – операторов, способствующих повышению уровня релевантности выдаваемых ссылок. В связи с развитием интеллекта поисковых машин возникла проблема правильного извлечения, обработки и понимания смысла естественного языка.

В работе рассматривается один из способов создания и функционирования мини-версии англоязычной интеллектуальной поисковой системы, а также возможные пути ее совершенствования. Под интеллектуальной поисковой системой понимается вопросно-ответная система, способная обрабатывать запросы пользователей, представленные в виде вопросительных высказываний, и находить на них точные ответы, опираясь на тексты базы данных поисковой системы. Экспериментальная система разработана на основе 19 текстов научно-популярных статей, извлеченных методом сплошной выборки из сайтов таких англоязычных научно-популярных журналов, как *National Geographic*, *Discovery* и др., а также 45 специальных вопросов по содержанию данных текстов, начинающихся с вопросительных слов *Where* и *When*. Вопросы были сформулированы жителями разных стран, для которых английский язык является родным, вторым или иностранным. Все вопросы составлены грамматически правильно; в них отсутствуют опечатки. Отобранные вопросительные предложения имеют простую синтаксическую структуру и небольшой размер. В них отсутствуют специальные символы, и они не содержат никаких знаков препинания, кроме вопросительного знака в конце предложения.

На основе материала исследования была смоделирована мини-версия англоязычной интеллектуальной поисковой системы, в основу работы которой положено достаточно объемное и четко структурированное лингвистическое обеспечение. Рассмотрим подробнее ее организацию.

Лингвистическая база данных англоязычной вопросно-ответной поисковой системы включает следующие компоненты:

1. Алфавитный тегированный словарь словоформ, входящих в запросы вопросительного типа и ответы на них, составленный с опорой на демоверсию POS-тегера Иллинойского университета, США.

2. Список вопросительных слов и дополнительных семантических тегов, соотнесенных с вопросительными словами.

3. Список последовательностей тегов, которые соответствуют словам, входящим в именные группы.

4. Список слов, распределенных по категориям и формирующих модули времени и места. По правилам, использованным при анализе вопросов, идущие подряд модули с одинаковым тегом были объединены, чтобы более полно отразить смысловую составляющую предложений.

5. Словарь контекстуальных синонимов, выделенных из проанализированных текстов и вопросов по их содержанию.

6. Список форм глаголов, использованных в запросах пользователей.

7. Список синтаксических структур вопросов и соответствующих им структур ответов.

В основу моделирования процесса функционирования англоязычной вопросно-ответной поисковой системы положено решение двух взаимосвязанных задач. Первая задача связана с проведением синтаксического анализа запроса пользователя, сформулированного в виде вопросительного предложения, и формированием его поискового образа (ПОЗ). Эта задача реализована в ходе построения формальной модели анализатора запросов. При решении второй задачи на основе ПОЗ осуществляется поиск предложений-кандидатов, содержащих необходимую пользователю информацию, извлечение этой информации и непосредственный синтез ответа. Данная задача реализована при разработке формальной модели синтезатора ответов.

Для проверки адекватности работы комплексной формальной модели был взят следующий текст англоязычной научно-популярной статьи:

The Japanese have just perfected the skateboard

And it looks amazingly fun.

A Japanese engineer just invented a nifty new way to travel: A transporter called a «WalkCar» that's small, light and apparently easy to use.

The product is battery powered and is about the size of a laptop. And although it looks like it can't hold much weight and is made from aluminum, it can apparently have as much as 265 lbs on board.

VentureBeat reported that it can go up to 6.2 miles per hour for up to 7.4 miles. It needs three hours to charge.

Creator Kuniako Saito told Reuters in an interview, «What if we could just carry our transportation in our bags, wouldn't that mean we'd always have our transportation with us to ride on?» and my friend asked me to make one, since I was doing my masters in engineering specifically on electric car motor control systems.»

Per VentureBeat:

Saito says customers will be able to reserve their own WalkCars from autumn 2015 on the crowd-funding website Kickstarter. The futuristic skateboard will have a price-tag of around 100,000 Japanese Yen (approx. $800 USD). Shipping is expected to begin by spring 2016.

Рассмотрим работу компьютера с данным текстом и несколькими специальными вопросами по его содержанию. В первой части формальной модели поступившие вопросы по содержанию текста подвергаются синтаксическому анализу. Например, это будут следующие вопросы:

1. *When will the shipping of WalkCars begin?*
2. *When can customers reserve WalkCars?*

На первом этапе анализа происходит деление предложений на токены для их последующего тегирования. Результат произведенных действий будет выглядеть следующим образом:

1. *When*/WRB/TM *will*/MD *the*/DT *shipping*/NN *of*/IN *WalkCars*/NNPS *begin*/VB.
2. *When*/WRB/TM *can*/MD *customers*/NNS *reserve*/VB *WalkCars*/NNPS.

На следующем этапе работы анализатора в системе создается два блока данных, в которых будут храниться структуры вопросов и их поисковые образы:

1.	**WRB**	**MD**	**DT NN**	**IN**	**NNPS**	**VBP**
	When	will	the shipping	of	WalkCars	begin
Блок А	WRB\TM	aux.v	Subj			MV
Блок Б	#MTM	the shipping of WalKCars will begin				

2.	**WRB**	**MD**	**NNS**	**VB**	**NNPS**
	When	can	customers	reserve	WalkCars
Блок А	WRB\TM	aux.v	Subj	MV	D.Obj
Блок Б	#MTM	customers can reserve WalkCars			

На этом работа анализатора запросов пользователей заканчивается, и данные отправляются во вторую часть формальной модели – синтезатор ответов. Данные из Блока Б помещаются в поисковую систему для поиска предложений-кандидатов. Как видно из текста, ни одного полностью совпадающего с запросом пользователя предложения нет, что приводит к необходимости изменить ПОЗ. Изменение и нормализация ПОЗ происходят следующим образом:

1. *#MTM the shipping of WalKCars will begin.*

2. *#MTM customers can reserve WalkCars.*

Сначала в поисковом образе запросов, с опорой на словарь синонимов заменяются некоторые слова. Поскольку контекстуальные синонимы есть только для второго запроса, первый запрос остается неизменным.

1. *#MTM the shipping of WalkCars will begin.*

2. *#MTM customers will be able reserve WalkCars.*

После этого поисковые образы запросов вновь направляются в поисковую машину. Поисковая машина найдет для второго запроса совпадение, потому что предложение-кандидат содержит многие слова из ПОЗ: *Saito says* ***customers will be able*** *to* ***reserve*** *their own* ***WalkCars*** **#MTM***from autumn 2015***#MTM** *on the crowd-funding website Kickstarter.* В то же время первый ПОЗ нуждается в корректировке. Поэтому следующим этапом изменения данного запроса является его упрощение, то есть удаление из него всех второстепенных по важности слов:

1. *#MTM shipping WalkCars begin.*

Поскольку прямое дополнение *WalkCar* встречается в тексте достаточно часто (что также будет отражено в поисковом образе документа), система делает вывод, что подобное слово может быть опущено:

1. *#MTM shipping begin.*

Используя такой запрос, система может найти подходящее предложение-кандидат ***Shipping*** *is expected to* ***begin*** **#MTM***by spring 2016****#MTM.*** Таким образом, после некоторых изменений поисковых образов запросов система может определить корректные предложения-кандидаты на ответ. Дальнейшая работа будет проводиться непосредственно с ними. При этом система должна опираться на знание частей речи, поэтому проводится тегирование найденных предложений-кандидатов.

1. *Saito*/**NNP** *says*/**VBZ** *customers*/**NNS** *will*/**MD** *be*/**VB** *able*/**JJ** *to*/**TO** *reserve*/**VB** *their*/**PRP$** *own*/**JJ** *WalkCars*/**NNPS** **#MTM***from*/**IN** *autumn*/**NN** *2015*/**CD#MTM** *onI*/***IN*** *the*/**DT** *crowd-funding*/**JJ** *website*/**NN** *Kickstarter*/**NNP**.

2. *Shipping*/**NN** *is*/**VBZ** *expected*/**VBN** *to*/**IN** *begin*/**VB** **#MTM***by*/**IN** *spring*/**NN** *2016*/**CD#*****MTM***.

Поскольку все необходимые для синтеза ответа данные получены, следующим этапом работы системы будет выбор структуры ответа, соответствующей запросу:

1. WRB/TM aux.v Subj Pred → Subj aux.v Pred #MTM.

2. WRB/TM aux.v Subj Pred D.Obj O.Comp → Subj aux.v Pred D.Obj O.Comp #MTM.

Далее, используя всю имеющуюся информацию, система заполнит выбранные структуры, в основном опираясь на вопросы пользователя:

1. ***The shipping of WalkCars will begin by spring 2016.***

2. ***Customers can reserve WalkCars from autumn 2015.***

Заключительный этап работы системы связан с окончательным лексико-грамматическим контролем, то есть удалением одинаковых слов и проверкой

правильного согласования частей речи. Как видно из приведенных примеров, использование при синтезе ответа слов из вопроса пользователя вместо синтеза предложения ответа с нуля позволяет повысить качество результатов работы системы.

Проверка адекватности работы формальной модели позволила сделать следующие выводы:

1. Компьютерная система, не проводящая глубинный семантический анализ и опирающаяся лишь на специализированные словари и ряд синтаксических правил, способна достаточно точно определить основной смысл поступившего вопроса пользователя, сформулированного на естественном языке.

2. При пополнении лингвистической базы данных компьютер может более точно выделять необходимую для поиска информацию.

3. Для улучшения работы системы необходимо проводить предобработку поступающих в нее текстов, что значительно сократит время работы поисковой машины.

4. При совершенствовании системы путем машинного обучения можно значительно сократить объем списка модулей и контекстуальных синонимов, что, в свою очередь, не только значительно уменьшит объем используемой системой памяти, но и повысит качество результатов работы.

Возможность применения полученных в ходе выполнения исследования результатов видится в использовании разработанной формальной модели в качестве основы для создания модуля промышленной интеллектуальной поисковой системы сети Интернет, способного обрабатывать запросы пользователей, представленные в виде вопросительных высказываний (общих и специальных вопросов), и находить на них точные ответы.

The article deals with the procedure of functioning and some ways of improvement of the mini-version of English question-and-answer retrieval system. The full formal model of the intelligent system consists of two parts – the question parser and the answer composer and functions on the basis of the complex linguistic knowledge base.