УДК 004.8

Д. В. Деревянко, **Д. Е. Пальчунов**

Новосибирский государственный университет
ул. Пирогова, *2*, *Новосибирск*, *630090*, *Россия*

Институт математики им. С. Л. Соболева СО РАН
пр. Акад. Коптюга, *4*, *Новосибирск*, *630090*, *Россия*

derevyanko.denis.v@gmail.com, *palch@math.nsc.ru*

ФОРМАЛЬНЫЕ МЕТОДЫ РАЗРАБОТКИ ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ [*]

Рассматриваются методы разработки вопросно-ответных систем. Особое внимание уделяется проблемам разработки вопросно-ответных систем с интерфейсом на естественном (русском) языке, порождающих ответы на вопросы на основе информации, содержащейся в документах, представленных в Интернете. Задача интеграции знаний, содержащихся в разных документах, тесно связана с проблемой порождения новых знаний, ни в одном тексте в явном виде не содержащихся. Описываются формальные методы автоматизированного извлечения знаний из текстов на русском языке и порождения по ним новых знаний, основанные на применении параметризованных запросов. На основе этих методов разработана первая версия вопросно-ответной системы.

Ключевые слова: вопросно-ответная система, информационно-поисковая система, извлечение знаний, порождение знаний, атомарная диаграмма модели, параметризованный запрос, анализ текстов естественного языка.

Введение

Статья посвящена проблемам и методам разработки вопросно-ответных систем. Задача разработки вопросно-ответных систем с интерфейсом на естественном языке – одна из наиболее актуальных задач в области информационного поиска [1–7] [1]. В настоящей работе мы рассматриваем проблему разработки вопросно-ответной системы с интерфейсом на русском языке.

Вопросно-ответные системы можно разделить на несколько типов: системы, порождающие ответ на вопрос пользователя при помощи своей локальной базы знаний; системы, порождающие ответ при помощи знаний, содержащихся в Интернете; и комбинированные системы. Мы рассмотрим задачу разработки вопросно-ответной системы второго типа: извлекающей знания из документов, написанных на естественном языке, представленных в Интернете. Такая система, по существу, является метапоисковой системой: для получения информации из Интернета мы пользуемся такими поисковыми системами, как Гугл и Яндекс.

[*] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 14-07-00903_а.

[1] *Manning Ch.* CS224N/Ling 284, Text-based Question Answering systems. 2013. URL: http://web.stanford.edu/class/cs224n/handouts/cs224n-QA-2013.pdf/; The START Natural Language Question Answering System, Computer Science and Artificial Intelligence Laboratory. 2014. URL: http://start.csail.mit.edu/index.php/; Evi official site, Our Technology. 2014. URL: https://www.evi.com/technology/; Evi (software). 2014. URL: http://en.wikipedia.org/wiki/Evi_%28software%29/.

Деревянко Д. В., *Пальчунов Д. Е.* Формальные методы разработки вопросно-ответной системы на естественном языке // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 3. С. 34–47.

Знания, необходимые для порождения ответа на вопрос пользователя, мы извлекаем из разных документов, представленных в Интернете. При этом возникает задача интеграции знаний, содержащихся в разных документах. Эта задача тесно связана с проблемой порождения новых знаний: комбинируя знания, извлеченные из разных документов, мы в итоге можем получить новые знания, ни в одном тексте в явном виде не содержащихся [8].

В работе описаны формальные методы автоматизированного извлечения знаний из текстов на русском языке и порождения по ним новых знаний. Эти методы основаны на формальном представлении знаний при помощи фрагментов атомарных диаграмм алгебраических систем [9]. Причем на языке фрагментов атомарных диаграмм мы формализуем как знания, извлеченные из текстов естественного языка, так и интегрированные, объединенные знания, на основе которых порождается ответ на вопрос пользователя.

Использование поисковых систем (Гугл, Яндекс) для извлечения знаний из текстов естественного языка, представленных в Интернете, основано на применении параметризованных запросов. При помощи параметризованных запросов мы также осуществляем интеграцию знаний, извлеченных из разных документов.

На основе описанных формальных методов разработана первая версия вопросно-ответной системы с интерфейсом на естественном языке.

Вопросно-ответные системы

Рассмотрим наиболее популярные вопросно-ответные системы.

START (аббревиатура от «SynTactic Analysis using Reversible Transformations») – первая в мире сетевая вопросно-ответная система на естественном языке, непрерывно работающая в онлайн-режиме с декабря 1993 г. Эта система была создана Борисом Кацем и его партнерами из лаборатории информатики и искусственного интеллекта Массачусетского технологического института. В отличие от информационно-поисковых систем целью системы START было снабдить пользователей «просто правильной информацией» вместо того, чтобы предоставить им список «релевантных» сайтов. В настоящее время система может ответить на миллионы вопросов на английском языке о местах (например, города, страны, озера, координаты, погода, карты, демография, политические и экономические системы), фильмах (например, названия, актеры, директора), людях (например, даты рождения, биографии), дать определения из словарей и многое другое.

В этой системе вопросы разбиты на четыре группы.

1. География.
   - Give me the states that border Colorado.
   - Show me a map of Denmark.
2. Искусство.
   - Who composed the opera Semiramide?
3. Наука и справочная информация.
4. История и культура.

Для хранения данных и выполнения запросов о фактах используется специальная «универсальная база» Omnibase. Она имеет модель «объект – свойство – значение», например, «Federico Fellini is a director of La Strada»:

- объект – «La Strada»;
- свойство – «director»;
- значение – «Federico Fellini».

Каждому объекту сопоставлен источник данных (data source), например: Star Wars – imdb-movie.

Система START является универсальной вопросно-ответной системой в отношении категорий обрабатываемых вопросов, при этом она использует только английский язык. Источниками знаний являются локальное хранилище (база знаний) и Интернет.

Другая известная вопросно-ответная система – Evi, которая доступна не только в виде веб-сайта, но и в виде мобильного приложения для различных платформ.

Evi (ранее называлась «True Knowledge») – технологическая компания в Кембридже, Англия, основанная Уильямом Танстол-Педо. Система осуществляет разбор предложенного во-

проса, снимая неоднозначность со всех возможных значений слов в вопросе, чтобы выявить наиболее вероятное значение ответа.

Накопление знаний и проверка. Evi пополняет информацию для своей базы данных двумя способами: импортирование из внешних источников данных (например, из Википедии) и сведения от пользователей. 21 ноября 2008 г. на официальном сайте компания объявила, что более чем 100 000 фактов были добавлены пользователями бета-версии. С августа 2010 г. база данных содержала 283 511 156 фактов о 9 237 091 объекте.

Рассмотрим также **Exactus** – поисково-аналитическую систему, разработанную в России [2].

Интеллектуальная метапоисковая система Exactus позволяет искать документы в сети Интернет. Поисковый запрос обрабатывается лингвистическими средствами Exactus и направляется к нескольким поисковым машинам (Яндекс, Рамблер, Гугл). Затем аннотации к документам обрабатываются лингвистическими средствами Exactus, сравниваются с запросом и выдаются пользователю. Низкорелевантные документы отбрасываются.

Поисковый алгоритм Exactus объединяет статистические и лингвистические методы поиска. Из статистических характеристик текста в Exactus учитываются TF*IDF веса термов (term frequency inverse document frequency – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса) и значимость фрагментов текстов. Оценка характерности слов документу рассчитывается на основе TF*IDF весов слов (термов). Далее, строятся прямой и обратный индексы слов, аналогично большинству поисковых систем.

Особенность Exactus заключается в том, что в индексах системы слова сортируются на основании их статистических характеристик с учетом смысловых значений слов. Иначе говоря, слова рассматриваются как синтаксемы. (Синтаксемой называется минимальная синтактико-семантическая единица языка, несущая свой обобщенный категориальный смысл в конструкциях разной степени сложности и характеризующаяся взаимодействием морфологических, семантических и функциональных признаков.)

Рис. 1. Пример работы системы Evi

[2] Интеграция лингвистических и статистических методов поиска в поисковой МАШИНЕ «Exactus». 2014. URL: http://www.dialog-21.ru/digests/dialog2008/materials/html/80.htm

В основу подхода, на котором основана система Exactus, положено следующее утверждение: смысл предложения определяется совокупностью входящих в него синтаксем и множеством связей между ними. Такой подход позволяет разделять слова с точки зрения лексики в различных семантических значениях в индексе (субъект, объект, результатив и т. д.). Это, в свою очередь, позволяет более тонко сопоставлять поисковый запрос и документы в индексе, находя только те документы, в которые входят слова в том же семантическом значении, что и в запросе. Таким образом, в результатах поиска документы, близкие запросу по смыслу, выдаются раньше остальных, что принципиально невозможно достичь при использовании только статистических методов.

На практике зачастую результаты, выдаваемые системой Exactus, отличаются от результатов поиска Google и Yandex не в лучшую сторону, особенно это заметно, если сравнить сниппеты, предоставляемые системами (рис. 2 и 3).

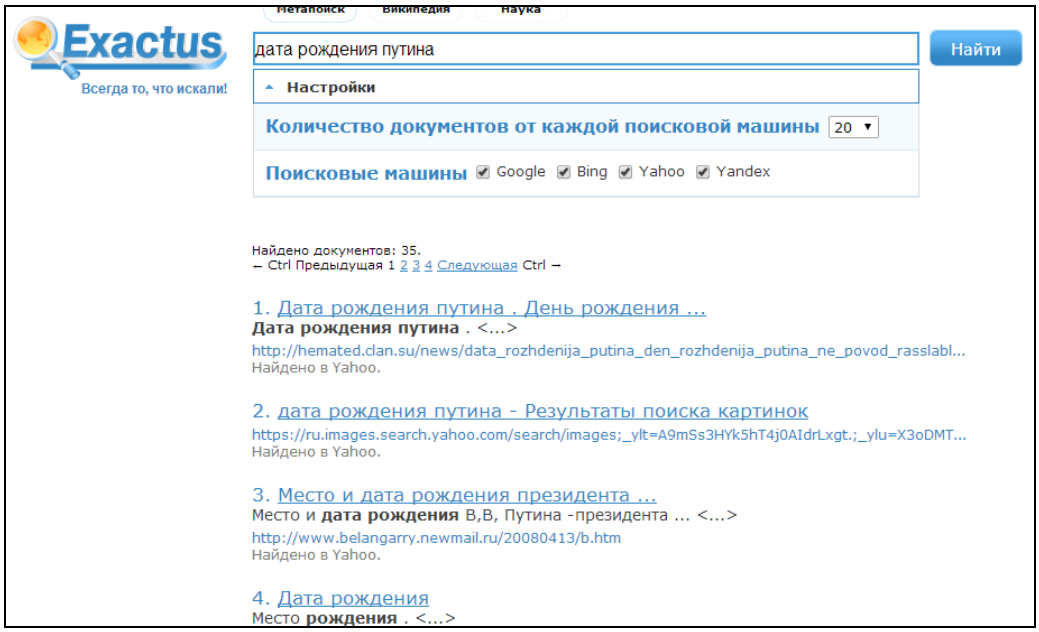

Рис. 2. Результаты запроса в категории «метапоиск»

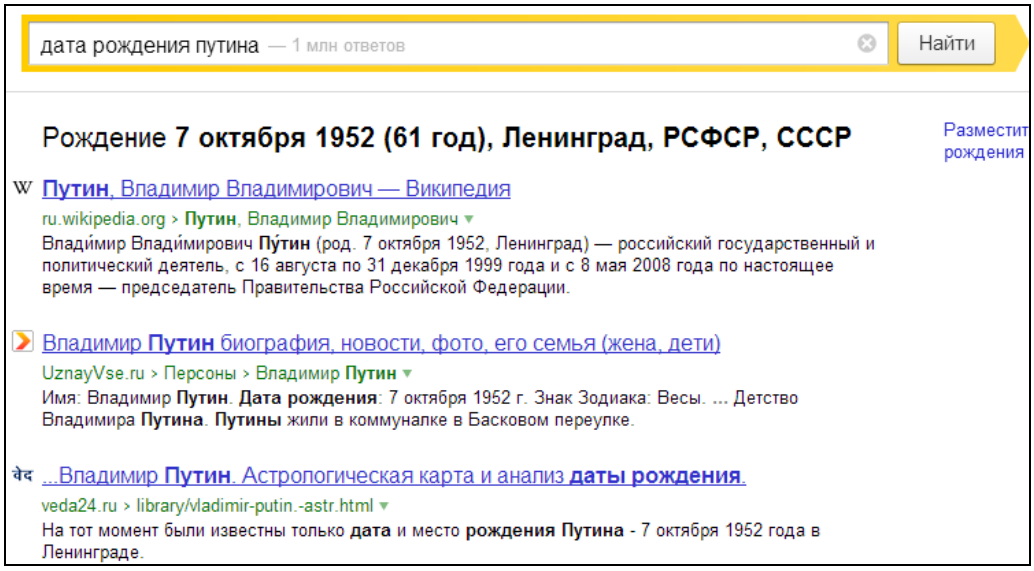

Рис. 3. Сравнение с результатами запроса в поисковой системе Яндекс

Формальное представление принципов работы вопросно-ответной системы на основе атомарных диаграмм моделей

Для решения задачи формального описания работы вопросно-ответной системы мы используем атомарные диаграммы алгебраических систем (моделей). Атомарным предложением называется бесквантворное предложение φ вида $P(c_1,...,c_n)$ – предикат от констант (здесь мы рассматриваем логику предикатов без равенства). Атомарной диаграммой модели называется множество истинных на ней атомарных предложений и отрицаний атомарных предложений.

Запросы пользователей, знания, извлеченные из документов, и ответы на запросы пользователей мы будем представлять в виде конечных множеств атомарных предложений и отрицаний атомарных предложений – конечных фрагментов атомарных диаграмм моделей. Каждый такой конечный фрагмент может быть представлен в виде одного предложения: конъюнкции $(\varphi_1 \& ... \& \varphi_n)$ набора атомарных предложений и отрицаний атомарных предложений $(\varphi_1,...,\varphi_n)$, составляющих данный фрагмент атомарной диаграммы модели.

Кратко опишем общий алгоритм работы разрабатываемой вопросно-ответной системы. Пользователь вводит запрос, представляющий собой одно или несколько предложений на естественном (русском) языке. При помощи технологии и программной системы, описанных в [9], по вопросу пользователя строится конечный фрагмент атомарной диаграммы модели. Сигнатура этого фрагмента сопоставляется с имеющимися онтологиями. В случае обнаружения омонимии (данное понятие – сигнатурный символ – входит в несколько онтологий и, следовательно, может иметь разный смысл) пользователю задается уточняющий вопрос. Например, «ключ – это родник, гаечный ключ, средство для открывания замка, скрипичный ключ… или ключ для шифрования?»; заметим, что указанный вопрос мы можем породить, воспользовавшись электронным тезаурусом WordNet [10] [3], который описывает не онтологию некоторой предметной области, а «верхнеуровневую» онтологию естественного языка. Далее при необходимости пользователю могут быть заданы уточняющие вопросы для разрешения анафор и кореференций.

В результате строится пополненный конечный фрагмент атомарной диаграммы модели, представляющий запрос пользователя на формальном логическом языке. Заметим, что при этом мы решаем проблему пертинентности информационного поиска в работе вопросно-ответной системы: пользователь получит ответ именно на тот вопрос, который он имеет в виду, а не просто информацию, синтаксически близкую к его вопросу. По существу, для этого мы решаем задачу извлечения умолчаний и пресуппозиций [11].

После формализации вопроса пользователя в виде конечного фрагмента атомарной диаграммы модели при помощи этого фрагмента порождается последовательность запросов к поисковой системе (Гуглу или Яндексу). Цель запросов – пополнение фрагмента атомарной диаграммы модели так, чтобы построенное множество предложений содержало необходимую пользователю информацию. Процесс пополнения атомарной диаграммы происходит итеративно: при порождении новых запросов к поисковым системам используется извлеченная ранее информация, представленная в новых предложениях, уже добавленных в атомарную диаграмму.

После этого по построенному фрагменту атомарной диаграммы модели порождается одно или несколько предложений естественного языка [9], являющиеся ответом на вопрос пользователя.

Один из наиболее важных способов пополнения фрагмента атомарной диаграммы модели для получения необходимой пользователю информации – заполнение пустующих мест («арностей») предикатов или, по терминологии теории «смысл –текст» [12; 13], валентностей терминов [9].

[3] WordNet – A lexical database for English. 2014. URL: http://wordnet.princeton.edu/; RussNet project. 2005. URL: http://project.phil.spbu.ru/RussNet/; Russicon Company Info. 2014. URL: http://www.russicon.ru/public.htm; Русский WordNet. 2014. URL: http://wordnet.ru/.

Рассмотрим в качестве очень простого примера вопрос «Где родился Пушкин?». Этот пример, конечно же, совершенно не показателен для демонстрации полезности вопросно-ответных систем: достаточно ввести в Гугл или Яндекс вопрос «Где родился Пушкин» и эти поисковые системы выдадут ответ. Дело в том, что для некоторых видов вопросов Гугл и Яндекс сами работают в качестве вопросно-ответных систем (но только для самых простых вопросов, как это будет видно из примеров ниже). Тем не менее в силу своей простоты этот пример хорошо подходит для иллюстрации используемых формальных методов.

Вопросу «Где родился Пушкин?» соответствует атомарное предложение *Родился_кто_где*(*Пушкин*, *x*), где *Родился_кто_где*(*a*, *b*) – двухместный предикат, *x* – неизвестная переменная и *Пушкин* – константа. Для того, чтобы ответить на этот вопрос пользователя, необходимо пополнить фрагмент атомарной диаграммы предложением *Родился_кто_где* (*Пушкин*, *Москва*), где *Москва* – константа, которая до этого могла как содержаться, так и не содержаться в исходном фрагменте атомарной диаграммы модели.

Для пополнения фрагмента атомарной диаграммы модели данным предложением *Родился_кто_где*(*Пушкин*, *Москва*) достаточно отправить в поисковую систему запрос «Пушкин родился в» и найти фрагмент текста «Пушкин родился в *C*», где *C* = *Москва* будет искомой константой; при этом нужно выделить фрагмент текста с константой требуемого типа: в данном случае место, но не дата.

Данное предложение и соответствующий запрос можно рассматривать как частные случаи предложений и запросов более общего вида: константу *Пушкин* можно считать значением некоторого параметра *Q*. Тогда предложение и запрос будут параметризованными: *Родился_кто_где*(*Q*, *x*) и «*Q* родился в *C*» соответственно. Для того чтобы получить конкретные предложение и запрос, нужно означить параметр *Q*.

Определение. Параметризованным атомарным предложением назовем предложение вида $P(c_1,...,c_n,Q_1,...,Q_l,x_1,...,x_k)$, где P – предикат, $c_1,...,c_n$ – константы, $Q_1,...,Q_l$ – параметры и $x_1,...,x_k$ – неизвестные переменные. Параметризованным запросом назовем запрос вида $t_1Q_1t_2...t_lQ_lt_{l+1}$, где $Q_1,...,Q_l$ – параметры, а $t_1,...,t_{l+1}$ – фрагменты текста.

Понятие параметризованного запроса – очень полезный инструмент при решении проблемы интеграции знаний, содержащихся в разных источниках [8; 9]. Покажем это на примере. Допустим, пользователь хочет получить ответ на следующий вопрос: «Где родился мэр самого большого города Сибири». Попробуем ввести такой запрос в Гугл. Ни в первых пяти результатах выдачи (рис. 4), ни в следующих нет даже намека на ответ на поставленный вопрос.

Заметим, что данный вопрос содержит несколько референтных индексов: указаний на неизвестные объекты, задаваемые (определяемые) своими свойствами. Это «самый большой город Сибири» и «мэр самого большого города Сибири». Имена этих объектов можно рассматривать как параметры Q_1 и Q_2, значения которых изначально неизвестны и, следовательно, их нужно определить. Поэтому целесообразно реализовать итеративную процедуру получения ответа на данный вопрос: определение одного за другим значений неизвестных параметров.

При этом, по существу, мы производим декомпозицию сложного поискового запроса на простые запросы. С помощью первого поискового запроса мы определяем значение первого параметра – выясняем название самого большого города Сибири. Для этого введем в Гугл запрос: «самый большой город Сибири» (в кавычках). Уже в пятом результате выдачи, в самом сниппете указано значение искомого параметра Q_1: «самый большой город Сибири – Новосибирск» (рис. 5).

Теперь для выяснения значения параметра Q_2 используем конструкцию параметрического запроса, определенную выше, а именно: параметрическим запросом в данном случае будет: «мэр Q_1». Подставив значение параметра Q_1, получаем запрос: «мэр Новосибирска» (рис. 6). Заметим, что при осуществлении данной синтаксической подстановки необходимо произвести денормализацию: заменить «Новосибирск» на «Новосибирска». Для этого нужно

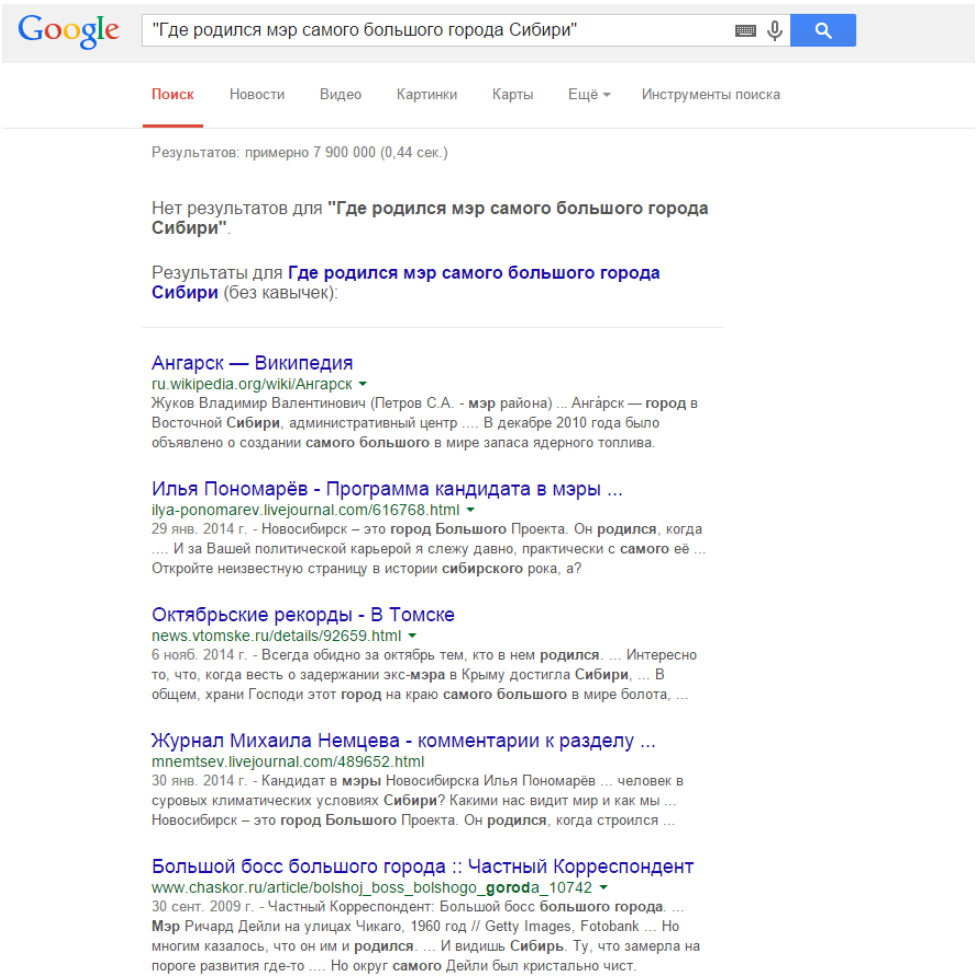

Рис. 4. Пример результатов поиска при прямом вводе вопроса

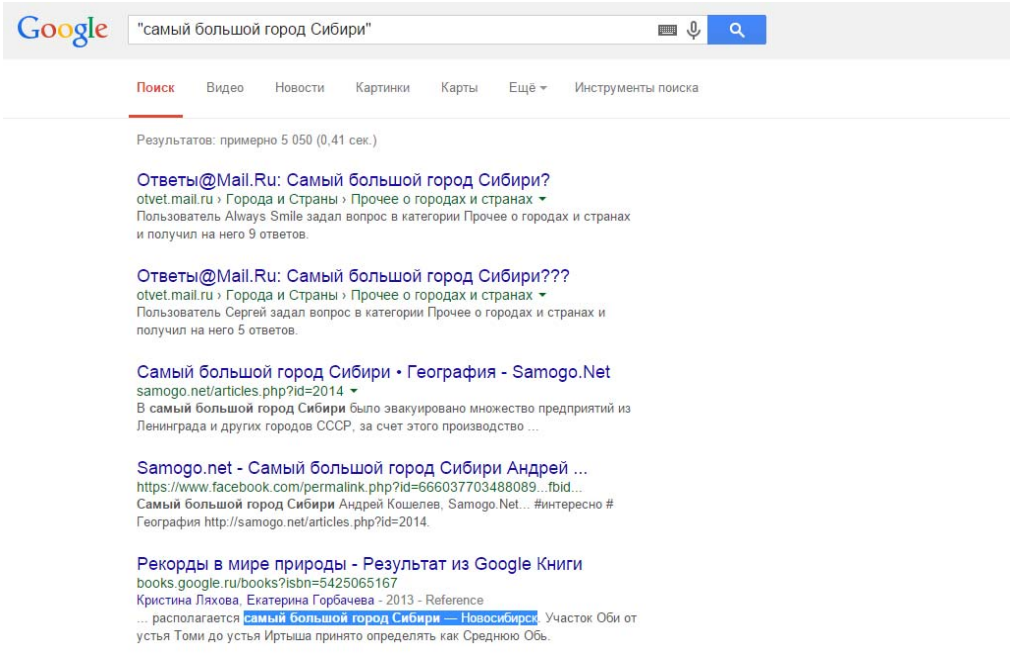

Рис. 5. Определение значения параметра Q_1

использовать словари нормализаций [4] и сочетаемости слов [14]. В пятом сниппете выдачи получаем: «Анатолий Локоть. Мэр Новосибирска». Эта информация, по существу, содержится и в первом сниппете: «Мэр Новосибирска высказался против принудительной эвакуации машин… Глава Новосибирска Анатолий Локоть высказался против принудительной эвакуации...», но не настолько явно: для извлечения значения искомого параметра Q_2 из первого сниппета необходимо при помощи онтологии отождествить два понятия: «мэр <Новосибирска>» и «глава <Новосибирска>».

Наконец, для получения ответа на исходный вопрос, снова воспользуемся параметрическим запросом: «Q_2 родился в». Подставив значение параметра Q_2, получаем запрос: «Анатолий Локоть родился в» (рис. 7). В первом же сниппете выдачи видим: «Анатолий Локоть родился в 1959 году в Новосибирске». Ответ на запрос «Анатолий Евгеньевич Локоть, Дата и место рождения» «дал» нам и сам Гугл, но несколько странным образом: «Анатолий Евгеньевич Локоть… Дата и место рождения: 18 января 1959 г. (55 лет)».

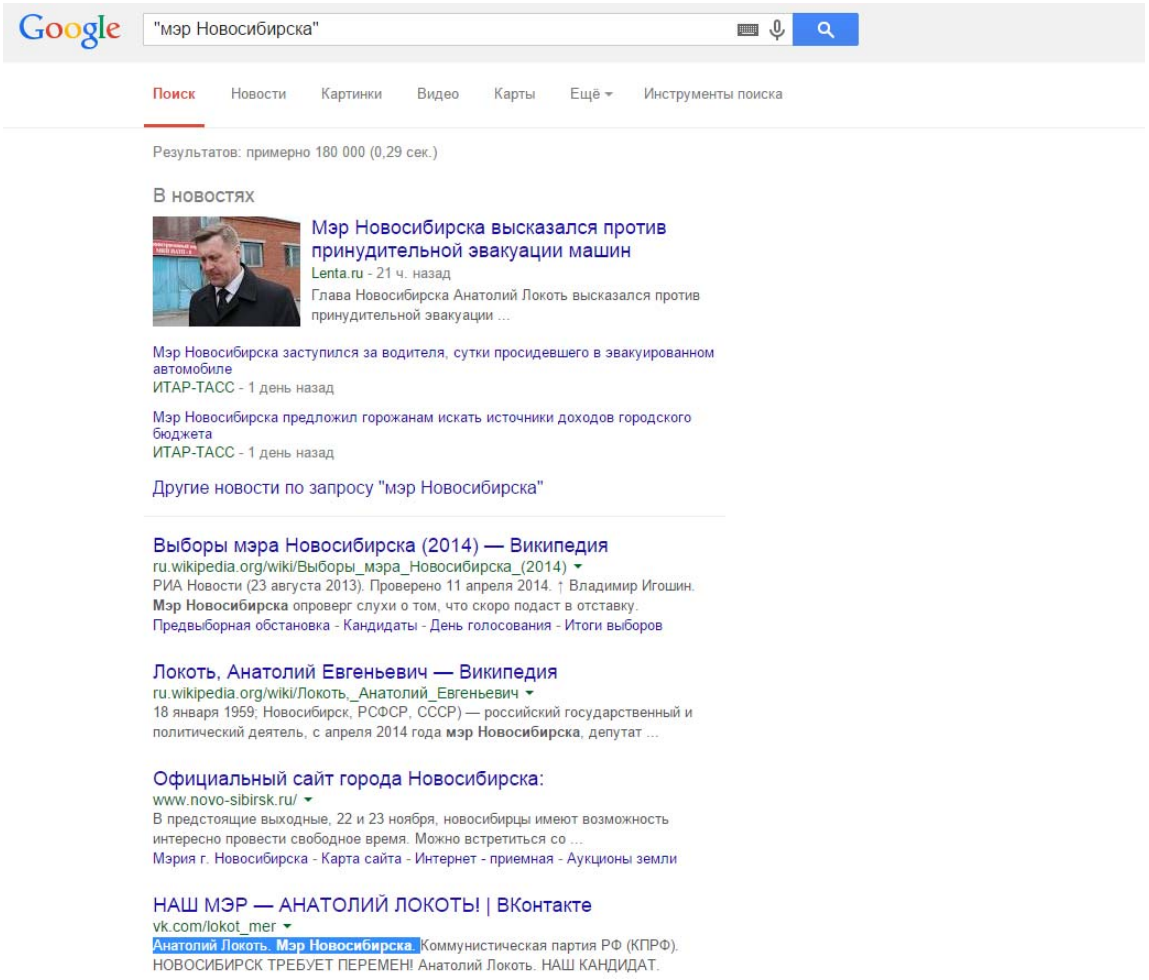

Рис. 6. Определение значения параметра Q_2

[4] Модуль морфологического анализа, проверки орфографии и лемматизации русского языка. 2014. URL: http://www.keva.ru/ling/rus/help.htm

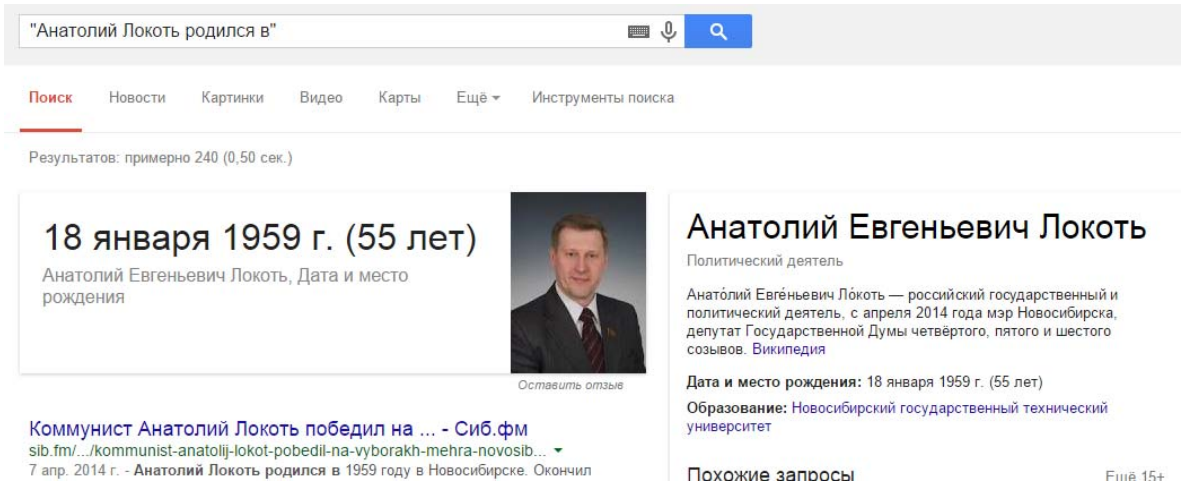

Рис. 7. Получение ответа на исходный вопрос

Рассмотренный пример является достаточно простым, но, несмотря на это, он показывает, что информация, в явном виде не содержащаяся в одном документе, может быть собрана из частей, содержащихся в нескольких документах. В определенном смысле это новая информация: если нет документа, в явном виде содержащего ответ на исходный вопрос. В этом случае возникает проблема объединения знаний, полученных из разных документов: интеграция разных онтологий и различных множеств пресуппозиций [11], разрешение омонимии [9] и др.

Первая версия вопросно-ответной системы на естественном языке

В настоящее время внимание специалистов по информационному поиску привлекают две важные проблемы [8]:

1) создание возможности формулировать поисковый запрос на естественном языке;

2) удовлетворение поисковой потребности вместо выполнения поиска документов, обладающих лишь определенными синтаксическими свойствами.

Для решения этих задач мы разрабатываем вопросно-ответную систему на естественном языке, поддерживающую поиск ответов на параметризированные запросы.

Созданная в ходе работы первая версия вопросно-ответной системы работает как метапоисковая система. Она обрабатывает запросы, сформулированные на русском языке, и выдает набор ответов с явным указанием на источники, откуда получена информация. В дальнейшем предполагается использовать как онтологии конкретных предметных областей, так и верхнеуровневые онтологии, а также WordNet и другие системы, в формальном виде представляющие семантику естественного языка.

Кратко опишем процесс работы первой версии вопросно-ответной системы, основываясь на принципах работы типовых вопросно-ответных систем [1; 2; 15]:

1. Анализ вопроса.
- Определение типа вопроса.
- Добавление синсетов к шаблонам вопроса.
- Построение дерева разбора, разметка частей речи, нахождение начальных форм для слов (нормализация).
- Выделение ключевых слов.

2. Информационный поиск.
- Формирование запросов к метапоисковой системе, порождение шаблонов для переформулирования вопросов.

- Получение и обработка результатов запроса, повторная фильтрация.
- Загрузка и фильтрация текстов, в которых потенциально содержится ответ; формирование единого списка всех фрагментов, упорядоченного по релевантности.

3. Извлечение ответа.

- Анализ фрагментов текста на совпадение с шаблонами ожидаемого ответа.
- Ранжирование фрагментов.
- Преобразование варианта ответа по шаблонам ответа для соответствия вопросу.

Приведенная схема может повторяться итерационно для определения и уточнения параметров запросов. В общем виде алгоритм состоит из следующих этапов.

- Анализ первоначального запроса пользователя.
- Выделение части запроса, описывающей параметр.
- Преобразование первоначального запроса.
- Замена параметра его конкретным значением.
- Информационный поиск при помощи поисковой системы (Гугла).
- Порождение ответа на исходный вопрос пользователя.

Остановимся на рассмотрении этапов более подробно.

Анализ запроса. На начальном этапе работы алгоритма происходит анализ вопроса. Пользователь вводит свой вопрос на русском языке. Введенный текст разбирается при помощи приложения CognitiveDwarf [16], которое производит синтаксический анализ текста и нормализацию используемых слов.

Далее определяется тип вопроса. Для этого при помощи регулярных выражений введенный пользователем текст классифицируется системой и с помощью шаблонов относится к одному из заранее определенных типов вопросов. В частности, определение типа вопроса происходит по содержащимся в нем вопросительным словам: *кто*, *что*, *какой*, *где*, *когда* и т. д. При этом используется ряд известных подходов к классификации вопросов [1].

По типу вопроса определяются шаблоны запросов к поисковой системе, эти шаблоны могут быть параметрическими. Значения параметров типизированы: например, в запросе о дате необходимо искать конкретные даты в различных форматах. Если нужно выяснить имя мэра Новосибирска, одним из шаблонов запроса к поисковой системе будет: «<имя человека> является мэром». Этот шаблон можно расширить с помощью синсетов – синонимических рядов тезауруса WordNet, и тогда он примет следующий вид: «<имя человека> [является|избран|назначен] мэром».

Синтаксический анализ вопроса, рассматриваемого как предложение русского языка, позволяет выделять и анализировать отдельные слова, входящие в предложение. Процесс выделения ключевых слов [17] реализуется следующим образом. При последовательном анализе предложения к набору ключевых слов добавляются [1]:

1) все слова и выражения, заключенные в кавычки;
2) все имена собственные;
3) все пары имя существительное – имя прилагательное;
4) все остальные существительные;
5) все глаголы;
6) фокус вопроса.

Фокус вопроса – это слово или последовательность слов, которые определяют вопрос и снимают неоднозначность: указывают на то, что ищется в данном вопросе, или на то, о чем этот вопрос. Например, для вопроса «Какой самый большой город в России?» его фокус – «самый большой город». Зная фокус и тип вопроса, легче определить тип ответа [16].

На основании синтаксического разбора строится дерево разбора вопроса, которое представляется в виде набора бинарных направленных отношений зависимости между словами в предложении. В каждом отношении задаются главное слово и зависимое слово. Тем самым получается дерево, которое в дальнейшем используется для реализации процедуры определения неизвестных сущностей (значений параметров в параметрических запросах).

Поиск информации, необходимой для построения ответа на вопрос пользователя, начинается с формирования запросов к поисковой системе. При работе программы к поисковой системе производится несколько запросов. Самым первым запросом к поисковой системе является сам вопрос пользователя, введенный полностью. Дело в том, что современные по-

исковые системы настолько развиты, что в определенных случаях мы можем сразу получить требуемый ответ на исходный вопрос пользователя.

Далее формируется последовательность поисковых запросов. Множество поисковых запросов расширяется за счет применения синсетов тезауруса WordNet. Например, из вопроса «Как защитить компьютер от атак вредоносных программ?» мы получаем другие вопросы:

- «Как обезопасить компьютер от атак вредоносных программ?»
- «Как предотвратить компьютер от атак вредоносных программ?»

На следующем шаге происходит информационный поиск при помощи поисковой системы Гугл. На этом этапе производится поиск релевантных документов, а также получение текстовых фрагментов, содержащих ответ.

Для того чтобы получить фрагменты текстов, которые будут содержать искомый ответ с наибольшей вероятностью, текст документа делится на части – абзацы. Выбирается фрагмент текста (абзац), набравший максимальное количество баллов; это количество баллов вычисляется при помощи специального алгоритма.

Извлечение информации, выбор фрагментов текста для порождения ответа пользователю. Стратегия отбора текстовых фрагментов, содержащих части ответа на вопрос, зависит от типа ожидаемого ответа. Например, для таких типов ответа, как географические местоположения, имена людей или дата события, при выборе релевантных фрагментов текста и извлечении частей ответа используются алгоритмы распознавания имен собственных, которые базируются на применении регулярных выражений.

Каждый фрагмент текста оценивается специальной функцией и выбирается фрагмент с максимальной оценкой. Оценка происходит, в частности, при помощи вычисления следующих величин [2]:

- keyword-score – количество ключевых слов, содержащихся в кандидате для ответа;
- keyword-distance-score – величина, равная расстоянию между кандидатом для ответа и ключевыми словами в тексте.

Также при оценке учитывается, встречаются ли слова из вопроса в кандидате для ответа в том же порядке, что и вопросе.

Программная реализация

С помощью первой версии разработанной нами вопросно-ответной системы пользователь может получать ответы на определенные типы вопросов, заданных на естественном языке. Пользователю может быть предложено несколько вариантов ответа на его вопрос (рис. 8), и он сам в таком случае должен выбрать наиболее подходящий ему ответ из списка, предложенного системой; при этом у пользователя есть возможность перейти на ресурс-первоисточник, чтобы убедиться в релевантности и достоверности найденной информации.

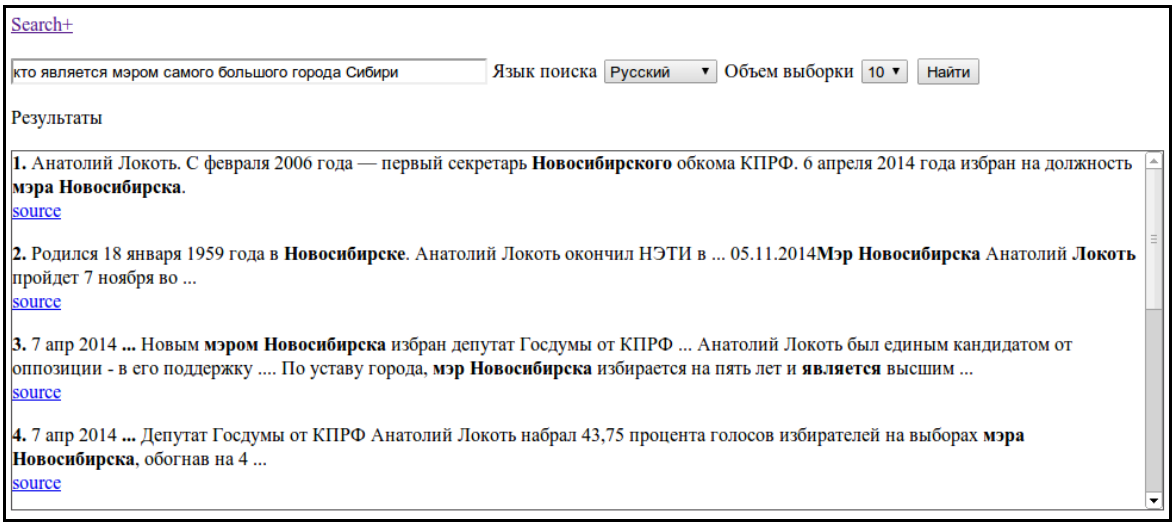

Рис. 8. Пример работы вопросно-ответной системы

В первой версии вопросно-ответной системы реализованы шаблоны для работы с именами людей и датами событий. Такие типы вопросов наиболее популярны [5]. Для обработки более сложных вопросов, например, вопросов о причинно-следственных связях, во-первых, нужно значительно большее число шаблонов и, во-вторых, необходимо улучшение алгоритмов извлечения знаний.

В разработанной вопросно-ответной системе используется поисковая система Гугл; без каких-либо проблем можно использовать также поисковую систему Яндекс. Язык вопросов и ответов с русского можно поменять на английский. В таком случае система будет включать в рассмотрение англоязычные ресурсы.

Заключение

Предложены формальные методы описания работы вопросно-ответной системы с интерфейсом на естественном языке, извлекающей необходимую для ответа информацию из текстов, представленных в Интернете.

Одним из путей порождения ответа на вопрос пользователя выступает анализ и интеграция знаний, содержащихся в разных текстах естественного языка. По существу, при таком объединении частичных знаний становится возможным порождение новых знаний, ни в одном из текстов в явном виде не содержащихся [8]. С другой стороны, если не известен документ, в котором это «новое» знание представлено, то это знание для нас действительно является новым (уже без кавычек).

При этом возникает проблема достоверности порожденного знания. В данной работе не ставится задача автоматического решения проблемы достоверности нового знания. Вместо этого пользователю предоставляются все первоисточники, и он сам может дать ответ: достоверна ли полученная информация или нет.

Разработанная программная система осуществляет поиск информации и предоставляет результаты, ранжированные по степени релевантности для пользователя. Вопросно-ответная система реализует разработанные методы и алгоритмы, в частности, предоставляет возможность получения ответов на параметризированные запросы.

Формализованы классы (типы) вопросов пользователей. Разработаны шаблоны, позволяющие определять тип вопроса, и шаблоны ответов на вопросы данных типов. В результате работы вопросно-ответной системы пользователь получает набор возможных ответов на свой вопрос и ссылки на тексты – первоисточники информации.

В дальнейшем предполагается использование онтологий предметных областей для получения ответов на вопросы по узкоспециализированным тематикам. Также предполагается реализовать методы извлечения и синхронизации множеств пресуппозиций [11], содержащихся в текстах естественного языка.

Список литературы

1. *Harabagiu S. M., Pasca M. A., Mariorano S. J.* Experiments with Open-Domain Textual Question Answering // Natural Language Engineering Journal. 2003. Vol. 9, iss. 3. P. 231–267.

2. *Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Girju R., Goodrum R., Rus V.* The Structure and Performance of an Open-Domain. Question Answering System // ACL '00 Proc. of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics Stroudsburg, 2000. P. 563–570.

3. *Hirschman L., Gaizauskas R.* Natural Language Question Answering. The View from Here // Natural Language Engineering. 2001. Vol. 7 (4). P. 275–300.

4. *Allam M. N., Haggag M. H.* The Question Answering Systems: A Survey // International Journal of Research and Reviews in Information Sciences. 2012. Vol. 2. No. 3. P. 211–221.

5. *Dwivedi S. K., Singh V.* Research and Reviews in Question Answering System // Procedia Technology. 2013. No. 10. P. 417–42.

[5] См.: *Manning Ch.* CS224N/Ling 284, Text-based Question Answering systems. 2013.

6. *Perera R., Nand P.* Interaction History Based Answer Formulation for Question Answering // Knowledge Engineering and the Semantic Web. Vol/ 468. Berlin; Heidelberg: Springer, 2014. P. 1–12.

7. *Chandrasekaran S., DiMascio C.* Create a natural language question answering system with IBM Watson and Bluemix services. 2014. URL: http://www.ibm.com/developerworks/cloud/library/cl-watson-films-bluemix-app/cl-watson-films-bluemix-app-pdf.pdf/.

8. *Пальчунов Д. Е.* Поиск и извлечение знаний: порождение новых знаний на основе анализа текстов естественного языка // Философия науки. 2009. № 4 (43). С. 70–90.

9. *Махасоева О. Г., Пальчунов Д. Е.* Автоматизированные методы построения атомарной диаграммы модели по тексту естественного языка // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 2. С. 64–73.

10. *Сухоногов А. М., Яблонский С. А.* Разработка русского WordNet // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. VI Всерос. науч. конф. (RCDL'2004). Пущино, 2004. С. 113–116.

11. *Пальчунов Д. Е., Целищев В. В.* Проблема извлечения знаний в системе взаимодействия человека и компьютера (онтологии и пресуппозиции) // Философия науки. 2012. № 4 (55). С. 20–35.

12. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл <=> Текст». 2-е изд., доп. М., 1999.

13. *Мельчук И. А., Жолковский А. К.* Толково-комбинаторный словарь современного русского языка. Вена, 1984.

14. Словарь сочетаемости слов русского языка / Под ред. П. Н. Денисова, В. В. Морковкина. 3-е изд., испр. М., 2002.

15. *Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Girju R., Rus V.* Lasso: A Tool for Surfing the Answer Net. National Institute of Standards and Technology (U.S.), 1999. P. 1–9.

16. *Антонова А. А.* Синтаксический анализатор для русского и английского языков // Информационно-аналитические аспекты в задачах управления: Сб. тр. ИСА РАН. М.: ЛКИ, 2007. Т. 29. С. 329–337.

17. *Пальчунов Д. Е., Степанов П. А.* Применение теоретико-модельных методов извлечения онтологических знаний в предметной области информационной безопасности // Программная инженерия. 2013. № 11. С. 8–16.

Материал поступил в редколлегию 24.11.2014

D. V. Derevyanko, **D. E. Palchunov**

Novosibirsk State University
2 Pirogov Str., *Novosibirsk*, *630090*, *Russian Federation*

Institute of Mathematics SB RAS
4 Koptyug Ave., *Novosibirsk*, *630090*, *Russian Federation*

derevyanko.denis.v@gmail.com, *palch@math.nsc.ru*

FORMAL METHODS OF DEVELOPMENT OF THE QUESTION-ANSWERING SYSTEM ON NATURAL LANGUAGE

The paper is devoted to the methods of development of question-answering systems. The special attention is given to the problems of development of question-answering systems having natural language (Russian) interface, which generate answers to the questions on the basis of information containing in the documents submitted in the Internet. The problem of integration of knowledge contained in different documents is closely connected with the problem of generation of new knowledge which is not explicitly presented in any text. Formal methods of automated extraction of knowledge from texts in Russian and generation of new knowledge, based on the use of the parame-

terized queries are described in the paper. On the base of these methods the first version of question-answering system has been developed.

Keywords: question-answering system, search engine, knowledge extraction, knowledge generation, atomic diagram of a model, parameterized query, analysis of natural language texts.

References

1. Harabagiu S. M., Pasca M. A., Mariorano S. J. Experiments with Open-Domain Textual Question Answering. *Natural Language Engineering Journal*, 2003, vol. 9, iss. 3, p. 231–267.

2. Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Girju R., Goodrum R., Rus V. The Structure and Performance of an Open-Domain. Question Answering System. *ACL '00 Proc. of the 38th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics Stroudsburg, 2000, p. 563–570.

3. Hirschman L., Gaizauskas R. Natural Language Question Answering. The View from Here. *Natural Language Engineering*, 2001, vol. 7, no. 4, p. 275–300.

4. Allam M. N., Haggag M. H. The Question Answering Systems: A Survey. *International Journal of Research and Reviews in Information Sciences*, 2012, vol. 2, no. 3, p. 211–221.

5. Dwivedi S. K., Singh V. Research and Reviews in Question Answering System. *Procedia Technology*, 2013, no. 10, p. 417–442.

6. Perera R., Nand P. Interaction History Based Answer Formulation for Question Answering. *Knowledge Engineering and the Semantic Web*, Berlin, Heidelberg, Springer, 2014, vol. 468, p. 1–12.

7. Chandrasekaran S., DiMascio C. Create a natural language question answering system with IBM Watson and Bluemix services. 2014. URL: http://www.ibm.com/developerworks/cloud/library/cl-watson-films-bluemix-app/cl-watson-films-bluemix-app-pdf.pdf/.

8. Palchunov D. E. Knowledge retrieval and elicitation: generation of new knowledge based on analysis of natural language texts. *Filosofiya nauki*, 2009, no. 4 (43), p. 70–90. (in Russ.)

9. Makhasoeva O. G., Palchunov D. E. Semi-automatic methods of a construction of the atomic diagrams from natural language texts. Vestnik of Novosibirsk State University, series: Information Technology, 2013, vol. 12, no. 2, p. 64–73. (in Russ.)

10. Sukhonogov A. M., Yablonsky C. A. Development of Russian WordNet. *Electronic libraries: perspective methods and technologies, electronic collections*. Proc. of the VI All-Russian scientific conference (RCDL'2004). Pushchino, 2004, p. 113–116. (in Russ.)

11. Pal'chunov D. E., Tselishchev V. V. The problem of knowledge retrieval in interaction between a man and a computer: ontologies and presuppositions. *Filosofiya nauki*, 2012, no. 4 (55), p. 20–35. (in Russ.)

12. Melchuk I. A. *Experience of the theory of the linguistic models «Semantic <=> Text»*. Moscow, 1999. (in Russ.)

13. Melchuk I. A., Zholkovsky A. K. *Sensible and combinatory dictionary of modern Russian*. Vienna, 1984. (in Russ.)

14. Denisov P. N., Morkovkin V. V. (eds.) The dictionary of word compatibility of Russian. Moscow, 2002. (in Russ.)

15. Moldovan D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Girju R., Rus V. *Lasso: A Tool for Surfing the Answer Net*. National Institute of Standards and Technology (U.S.), 1999, p. 1–9.

16. Antonova A. A. Syntax analyzer for the Russian and English languages. *Information and analytical aspects in problems of management*. Moscow, LKI, 2007, vol. 29, p. 329–337. (in Russ.)

17. Palchunov D. E., Stepanov P. A. The use of model-theoretic methods for extracting ontological knowledge in the domain of information security. *Programnaya ingeneriya*, 2013, no. 11, p. 8–16. (in Russ.)