

УДК 004.891.2

DOI: 10.15827/2311-6749.16.4.1

ФОРМИРОВАНИЕ ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЫ В УСЛОВИЯХ ОГРАНИЧЕННОГО ОБЪЕМА СЕМАНТИЧЕСКИ РАЗМЕЧЕННОГО КОРПУСА

*С.А. Беляев, к.т.н., доцент, bserge@bk.ru; А.С. Кулешов, студент, yol0317@gmail.com
(Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина) (СПбГЭТУ «ЛЭТИ»), ул. Профессора Попова, 5, г. Санкт-Петербург, 197376, Россия)*

В статье рассмотрены существующие подходы к реализации вопросно-ответных систем, готовые решения, которые могли бы использоваться в качестве основы, выявлены их достоинства и недостатки, предложены альтернативные подходы к построению системы, представлена их функциональная структура. Предложена математическая модель решения задачи, приведены результаты экспериментов в условиях ограниченного объема семантически размеченного корпуса для сравнения качества работы альтернативных решений.

Ключевые слова: *вопросно-ответная система, QAS, семантико-синтаксический анализатор, SRL, семантически размеченный корпус.*

Современные поисковые системы обеспечивают выдачу множества разных типов информации в виде веб-страниц, документов, изображений, видео, новостей и карт, но до сих пор не в состоянии в полной мере распознавать запросы на естественном языке и формировать ответы в соответствующем формате. Подобных запросов, согласно проведенным поисковой системой Яндекс исследованиям [1], задается порядка полутора миллионов в день, что составляет более 1 % от общего ежедневного потока запросов. Как результат подобных запросов – список релевантных ссылок, по которым пользователю предстоит осуществить дополнительный поиск информации, что является неоптимальным вариантом с точки зрения затраченных пользователем времени и ресурсов. В связи с этим в последнее время наблюдается смещение акцентов в сторону использования интеллектуальных систем поиска информации, каковыми считаются вопросно-ответные системы.

Вопросно-ответная система (QA) в общем случае представляет собой информационную систему, аккумулирующую в себе комплекс справочных и интеллектуальных систем, использующих естественно-языковой интерфейс. На вход QA-системе формируется вопрос на естественном языке, обработав который, система генерирует естественно-языковой ответ. В качестве источников данных система использует как локальные хранилища, так и глобальную сеть Интернет.

В настоящий момент известные реализации вопросно-ответных систем имеют показатель качества работы ниже 60 %. Оценка качества работы таких систем осуществляется на тестовых данных, разработанных обществом экспертов и представленных в виде фрагмента текста, вопроса по этому тексту и предполагаемого ответа. В случае, если система по запросу выдает результат, не соответствующий предполагаемому ответу, результат не засчитывается.

В 2012 году проводились соревнования по качеству работы вопросно-ответных систем в области биомедицины, наилучший результат (с точностью 55 %) показала система [2]. Отечественных систем, способных составить конкуренцию зарубежным системам по качеству работы, на данный момент не существует.

Проблема усугубляется тем, что как таковая задача установки семантических ролей для русского языка не поставлена в отличие от английского, где подготовлена существенная база аннотированных текстов, как следствие – отсутствие полноценного корпуса с семантической разметкой. Поэтому решение поставленных задач должно подтолкнуть к развитию отечественных подходов к автоматической обработке текста и увеличить объемы существующих размеченных корпусов.

Предлагаемая научно-инновационная идея может стать основой для широкого спектра научных исследований, которые позволят повысить интеллектуальность современных поисковых систем и сделать очередной шаг к полноценному искусственному интеллекту.

Разработанная система может применяться как при поиске необходимых новостей, так и для анализа текстов по контексту упоминания различных фактов в сети Интернет, например, для анализа преднамеренных информационных трендов и выяснения их источников. Учитывая сложную внешнеполитическую обстановку, даже частичное решение задач автоматического анализа текстов позволит определять тенденции по изменению настроения в обществе без проведения дорогостоящих социологических опросов.

Таким образом, можно сделать вывод, что данная задача имеет большое прикладное значение.

Современные QA-системы построены на принципах NLP (Natural Language Processing), что позволяет им образовывать классы интеллектуальных систем информационного поиска. Системы имеют как общие

принципы обработки естественного языка, так и различия в способе внутреннего представления семантического содержимого и в алгоритмах для приведения текста на естественном языке к внутреннему представлению системы. Следовательно, результаты работы системы напрямую зависят от реализуемых алгоритмов компьютерной лингвистики.

START (SynTactic Analysis using Reversible Transformations) [3]. Данная система является первой в своем роде QA-системой на естественном языке, была запущена в конце 1993 года. Работает в онлайн режиме. Ее главная цель – предоставить пользователю готовый ответ вместо того, чтобы предложить список релевантных сайтов. Система поддерживает лишь английский язык, отвечает на вопросы, предварительно распределяя их по категориям:

- наука и справочная информация;
- искусство;
- география;
- история и культура.

Для своего функционирования система использует такие техники, как аннотирование естественного языка, тернарное представление, правила трансформации, а также совместное накопление знаний. В процессе работы происходит обучение системы посредством накопления знаний. Кроме того, имеется возможность расширения базы знаний за счет пользователей.

EVI (прежнее название – True Knowledge) [4]. Особенность этой системы такова, что любой задаваемый вопрос разбирается для снятия полной неоднозначности со всех возможных слов, составляющих вопрос. Данный подход позволяет наиболее точно выявить вероятное значение ответа.

Пополнение БД системы осуществляется двумя способами: импорт из внешних источников данных, таких как сеть Интернет, а также добавление сведений от пользователей. БД системы содержит около 300 млн фактов примерно о 9 млн объектов.

AQUA – вопросно-ответная система, представляющая собой консольное приложение, работающее в Unix-подобном окружении [5]. Для перевода текста на естественном языке во внутреннее представление используется NLP-анализатор, осуществляющий разбиение фраз на лексемы. В качестве внутреннего представления выбран язык логических запросов, представляющий подмножество языка Prolog.

Одной из главных задач данной системы является классификация вопроса. Формат вывода ответа непосредственно зависит от типа вопроса (например: что, где, когда, который, почему, зачем и т.д.).

Система обеспечивает и полуавтоматический анализ ошибок. В случаях, когда пользователь получает заведомо ложный результат, у него есть возможность указать на ошибку и предложить свой вариант ответа, который впоследствии может использоваться системой.

Answers.com – система, генерирующая ответы по определенному признаку, термину или персоне [6]. Признак, по которому формируется ответ, составляется по типу запроса, например, «что такое», «кто такой». Также в системе предусмотрен вариант получения ответа от эксперта. Answers.com представляет собой интерактивный портал, где каждый пользователь может выступать как в роли эксперта, так и в роли задающего вопросы.

WolframAlpha – QA-система, разработанная в 2009 году [7]. Данная система использует математический пакет Mathematica для решения различных математических задач, таких как вычисление интегралов, построение графиков функций и многое другое. WolframAlpha использует механизм предоставления ответа на основе имеющихся знаний. Система поддерживает запросы как на естественном языке, так и специально сформированные в соответствии с внутренним стандартом.

Exactus – отечественная разработка вопросно-ответной системы [8]. Представляет собой интеллектуальную метапоисковую систему, которая позволяет искать документы в сети Интернет.

Механизм работы системы инициируется пользовательским запросом, который обрабатывается и перенаправляется на несколько поисковых машин (Google, Яндекс, Bing, Yahoo). Далее выполняется аннотирование документов посредством лингвистических инструментов Exactus, по окончании наиболее релевантные документы выдаются пользователю.

Алгоритм работы системы объединяет в себе статистические и лингвистические методы поиска, в качестве статистических характеристик текста используется TF*IDF (term frequency inverse document frequency).

Рассмотренные аналоги QA-систем нельзя применить для поставленной задачи – поддержка русского языка, исключением является система Exactus, основанная на статистических алгоритмах обработки текста, эффективность такой системы видится недостаточной, ввиду чего актуальным решением является разработка вопросно-ответной системы, основывающейся на развитой лингвистической теории семантики, которая должна повысить эффективность решения многих задач.

В качестве возможного подхода к разработке вопросно-ответной системы рассмотрим вариант, базирующийся на семантическом анализаторе русского языка. В виде семантического анализатора выступает SRL (семантическая разметка ролей текста). Этот подход представляет собой лингвистический метод семантического разбора. Метод относится к обобщению. При обобщении для подготовки аннотации ис-

пользуются системы обработки естественных языков (NLP), такие как грамматики и словари для синтаксического разбора, генерация естественно-языковых конструкций и онтологические справочники.

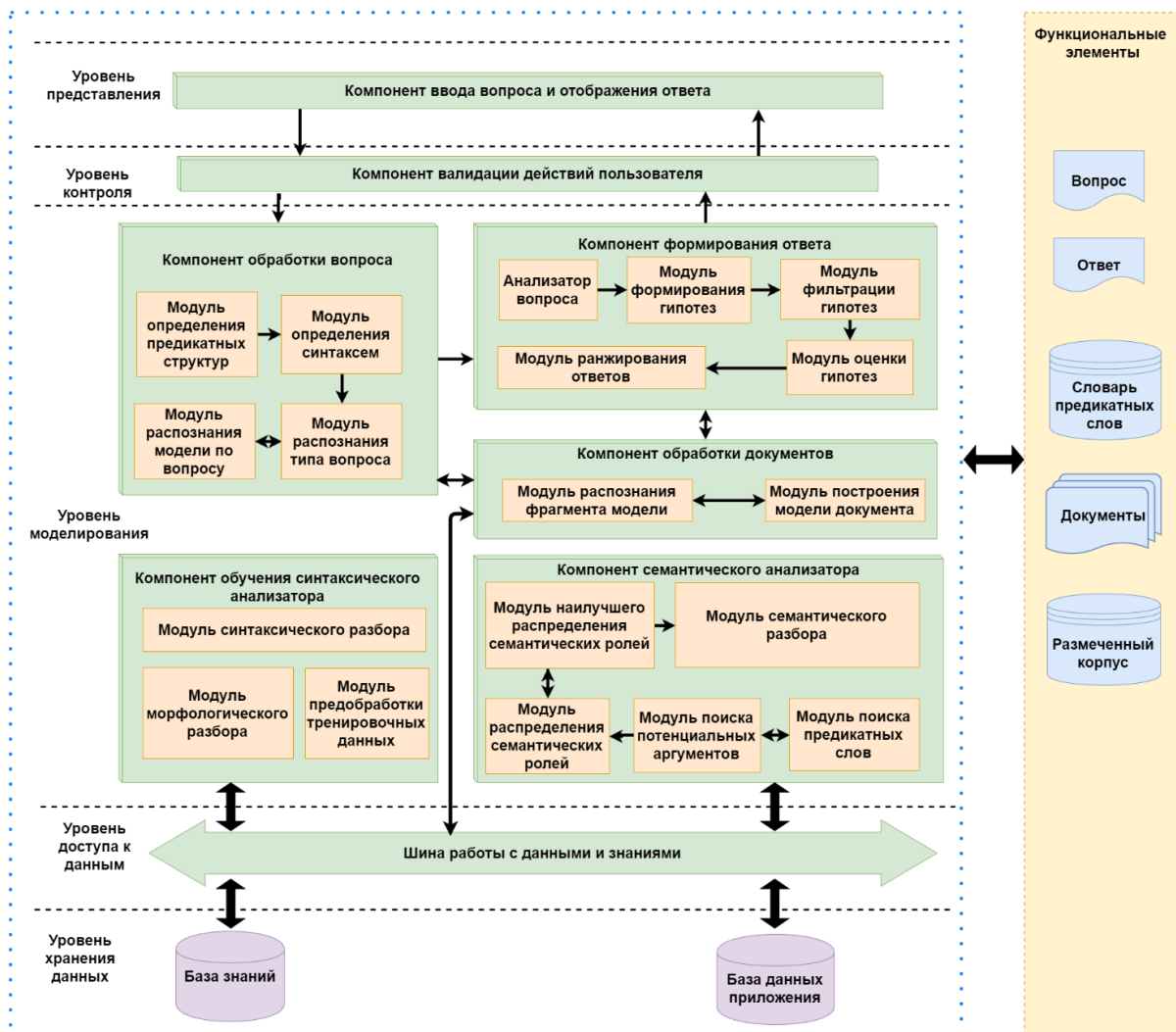
Центральным и самым трудоемким элементом этой системы является семантический анализатор. Качество работы семантического анализатора значительно повышается, если во время разметки использовать информацию, представленную синтаксическим анализатором по этому тексту. Проблемой является то, что использование этапа синтаксического анализа значительно сказывается на производительности системы. Наилучший из алгоритмов, решающих эту задачу, обладает асимптотической сложностью $O(n)$.

Предлагается рассмотреть и экспериментально сравнить несколько альтернатив, позволяющих решить поставленную задачу:

- семантический разбор источника данных и задаваемого вопроса;
- семантико-синтаксический разбор источника данных и задаваемого вопроса;
- синтаксическая связь как семантическое отношение.

Помимо семантического анализа и его модификации с дополнительным этапом синтаксического разбора, рассмотрим еще один альтернативный подход [9], представляющий видоизмененный вариант семантико-синтаксического разбора, в котором процессы синтаксического и семантического разборов происходят параллельно, синтаксическим связям ставятся в соответствие метки семантических отношений. Такая неразрывная связь семантического и синтаксического анализа гарантирует повышение качества работы системы. Следовательно, назначая семантические значения в качестве метки, результатом работы анализатора будут совместно построенное синтаксическое дерево предложений и его семантическая структура. Недостатком такого подхода является невозможность в частных случаях однозначно восстановить семантические связи между синтаксемой и предикатным словом. Частный случай обусловлен тем, что не все синтаксемы синтаксически связаны с предикатами [10].

Общий процесс работы вопросно-ответной системы в соответствии с представленной на рисунке архитектурой заключается в последовательном выполнении следующих трех стадий.



Архитектура вопросно-ответной системы

I. Синтаксический анализ. На начальном этапе выполняется обучение классификатора синтаксического анализатора. Задача анализатора – предсказывать переходы (последующие действия) по синтаксическому дереву зависимостей.

Для обучения классификатора синтаксического анализатора используется синтаксический размеченный корпус. Данные корпуса разделяются на тренировочные и тестовые выборки, в качестве единицы разбиения обычно используется предложение.

Тренировочная выборка осуществляет обучение классификатора, а тестовая используется для оценки точности работы классификатора. Обе выборки поступают на вход модуля извлечения свойств, который преобразует информацию в свойства, используемые классификатором, каждому узлу синтаксического дерева зависимостей приписывается набор признаков, в том числе метка класса. Для тренировочной выборки метка класса используется в качестве обучения классификатора, для тестовой – в целях сравнения результатов работы системы с эталонной разметкой.

После окончания подготовительного этапа обучения классификатора выполняется синтаксический разбор текста. Существует множество реализаций детерминированных алгоритмов синтаксического разбора текста [11], тем не менее, можно выделить явного лидера, обладающего достаточно высокой точностью и наименьшей асимптотической сложностью, – алгоритм Nivre.

Алгоритм представляет последовательную, слева направо, обработку слов в предложениях. По имеющимся признакам, таким как слово, контекст слова и частично построенное синтаксическое дерево, классификатор прогнозирует следующие действия анализатора. В алгоритме используются четыре операции, взаимодействующие со стеком:

- пропуск текущего правого слова;
- выталкивание одного слова из стека;
- проведение левой связи (устанавливается зависимость между входным словом и словом на вершине стека, затем выполняется выталкивание одного слова из стека);
- проведение правой связи (устанавливается зависимость между словом на вершине стека и входным словом, затем входное слово кладется на верхушку стека).

II. Семантический анализ. Случай, когда системой выполняется синтаксический разбор, результат его работы поступает на вход в семантический анализатор. Далее выполняется сопоставление синтаксического дерева с разметкой по семантическим ролям. В итоге схема семантической разметки будет ограничиваться лишь синтаксическими узлами при семантико-синтаксическом анализе и фрагментами текста при семантическом анализе. Для каждой семантической роли в схеме ставится в соответствие узел дерева зависимости.

Семантический анализ описывается последовательностью выполнения следующих шагов:

- поиск предикатных слов и соответствующих им словарных статей в словаре предикатных слов;
- поиск потенциальных аргументов предикатных слов;
- назначение семантических ролей найденным аргументам в соответствии с каждой найденной словарной статьей;
- выбор наилучшего распределения семантических ролей по аргументам для заданной словарной статьи с решением учета ограничения единственности ролей;
- определение наилучшей словарной статьи для предикатных слов и соответствующего ей набора семантических ролей.

III. Поиск ответа на вопрос. Этот этап включает в себя целый комплекс задач:

- анализ вопроса;
- формирование гипотез;
- фильтрация;
- оценка гипотез;
- ранжирование ответов.

Задача анализа заключается в установлении категории вопроса, которая задается с помощью модуля классификации. На основе этой информации выбирается соответствующий алгоритм поиска ответа и определяется источник информации. Для классификации вопроса используется онтология, предложенная А. Грессером [12]. Данная онтология описывает 18 категорий: контроль, дизъюнкция, дополнение, свойство, квантификация, определение, пример, сравнение, интерпретация, причина, последствие, цель, процедура, возможность, ожидание, суждение, отсутствие, запрос. Одновременно с классификацией выполняются определение фокуса вопроса и его семантический разбор. Определение фокуса представляет собой выявление некоторой части вопроса, которую, заменив ответом, можно представить в виде полноценного предложения.

На этапе формирования гипотез осуществляется поиск информации во всевозможных источниках, таких как газеты, книги, сеть Интернет.

После построения множества гипотез, которые выступают в качестве ответов, осуществляется фильтрация, которая отсеивает по определенным свойствам большинство кандидатов в ответы. Главная цель фильтрации – сократить время выполнения последующего этапа оценки гипотез.

Этап оценки задействует множество ресурсозатратных алгоритмов. Оценка заключается в подтверждении или опровержении факта гипотезы, в это время осуществляются запросы к различным источникам данных. В случае подтверждения формируется качественная оценка гипотезы.

На заключительном этапе осуществляется ранжирование ответов, то есть окончательная оценка достоверности каждого ответа для заданного вопроса и определение позиции в соответствии с его качественной оценкой.

Предлагается сформулировать модель решения задачи поиска ответа на вопрос в виде картежа $A = \langle SI, SE, D_\tau, Q \rangle$, где D_τ – массив текстов ($\tau = \overline{1, R}$), R – общее количество текстов, представленных источником данных; SI – модель синтаксического анализатора, описываемая следующим образом:

– $T_x = \{t_{xij}\}$ – набор тренировочных текстов с синтаксической разметкой в формате XML ($i = \overline{1, n}$, $j = \overline{1, m}$, n – количество слов в j -м предложении, m – количество предложений);

– $CONV_{SI}$ – конвертер, выполняющий отображение тренировочных текстов в формате XML в формат MALT-TAB. Представляет собой однозначно определенную функцию $CONV_{SI}: T_x \rightarrow T_m$, которая ставит в соответствие каждому элементу $t_x \in T_x$ единственное значение $CONV_{SI}(t_x)$, принадлежащее множеству T_m . Каждый элемент t_x обладает набором свойств $t_x = \{word, feat, dom, link\}$, производным свойствам $t_m = \{form, posttag, head, deprel\}$, где $form$ – форма слова, $posttag$ – «частеречная» разметка, $head$ – синтаксически главное слово (в виде id слова), $deprel$ – отношение зависимости к главному слову.

– S_{SI} – классификатор, формирует обученную модель на основе метода опорных векторов, идея которого – построить оптимальную разделяющую гиперплоскость на основе тренировочных данных, максимально удаленную от экземпляров разных множеств, наиболее близких к границе класса.

Гиперплоскость в признаковом пространстве описывается формулой вида $vx + z = 0$, где v – вектор, перпендикулярный гиперплоскости, z – постоянная, указывающая расстояние гиперплоскости до центра координат. Тогда для пары классов А и В положительное значение целевой функции гиперплоскости $f(x) = vx + z$, определяет экземпляр x к классу А, отрицательное – к классу В. Задача максимизации удаленности гиперплоскости от ближайших экземпляров разных классов заключается в минимизации величины $\|v\|$ имеющей место в формуле описания расстояния от гиперплоскости до экземпляра

$x_i: d((v, z), x_i) = \frac{y(x_i v + z)}{v} \geq \frac{1}{v}$. Каждый экземпляр класса характеризуется следующим набором свойств:

$x_i = \{word, pos, prefix, postfix, list_of_pos(x_1..x_{i-1})\}$, где $word$ – слово, pos – часть речи, $prefix$ – приставка, $postfix$ – окончание, $list_of_pos$ – набор частей речи предшествующих слов предложения.

Синтаксический разбор вопроса (Q) и источника данных (D_τ) осуществляется посредством задействования модели стека, поддерживающей четыре основные функции:

- пропуск текущего правого слова;
- выталкивание одного слова из стека;
- проведение левой связи, таким образом устанавливается зависимость между входным слово и словом на вершине стека, затем выполняется выталкивание одного слова из стека;
- проведение правой связи, таким образом устанавливается зависимость между словом на вершине стека и входным словом, затем кладется входное слово на верхушку стека.

SE – модель семантического анализатора, описывается следующим образом:

– $V = \{w_{ij}\}$ – словарь предикатных слов ($i = \overline{1, n}$, $j = \overline{1, m}$, m – количество словарных статей для i -го предикатного слова, n – количество предикатных слов в словаре);

– S_{SE} – классификатор, осуществляющий поиск предикатных слов предложения, их аргументов и семантических ролей, соответствующих аргументам, таким образом: $\forall w_p, \exists w_v \in V \Rightarrow args(w_p) \subseteq args(w_v)$.

Тогда $\forall j: \sum_i w_{p_{ij}} = 1; \forall i \neq \emptyset: \sum_j w_{p_{ij}} \leq 1$, где j – семантическая роль i -го предиката очередного предложения.

Процесс формирования ответа на вопрос проходит множество этапов.

– классификация вопроса (Q): $(\exists q \in Q) S_Q(q) \in C$, $C = \{cas, goal, enab, verf, disj, proc, conc, exp, jud, qua, feat, reg, def, ex, intr, assr, comp\}$;

– фокус вопроса: $F: Q \rightarrow A^*$, где $q = \{w_i\}$, $q \in Q$; $a = \{w_j\}$, $a \in A^*$, $q \supseteq a$;

– построение множества гипотез, описывается функцией $SEEK: A^* \rightarrow H$, где H – множество гипотез;

– выбор гипотез для заключительного этапа формирования ответа, формируется множеством $\{h \in H \mid h = true\}$, где $true$ – подтвержденный факт гипотезы h ;

– ранжирование ответов: $sort(h \in H, eval(h))$.

Для оценки качества работы вопросно-ответной системы, базирующейся на разных альтернативах (описанных ранее), были проведены экспериментальные исследования. В целях решения поставленной задачи использован специализированный набор тестов РОМИП [13], включающий в себя порядка 300 вопросов на русском языке о некоторых фактах вместе с готовыми ответами, покрывающими всевозможные альтернативы. В ходе работы был проведен эксперимент, в рамках которого получены результаты (см. таблицу), характеризующие оценку качества для всех трех альтернативных реализаций вопросно-ответной системы, базирующихся на разных подходах семантического разбора.

Результаты качества работы разработанных вопросно-ответных систем

Альтернатива	Всего вопросов	Отвеченные вопросы	Корректно отвеченные	Точность	Отзывчивость
Семантический подход	326	112	66	0,51	0,18
Семантико-синтаксический подход		283	136	0,42	0,37
Описание синтаксических связей как семантических отношений		105	61	0,51	0,17

Графа «Отвеченные вопросы» включает число вопросов, для которых соответствующий модуль QA-системы давал как минимум один ответ. «Корректные ответы» – количество вопросов, на которые был получен правильный, наиболее релевантный ответ. «Точность» и «Отзывчивость» были рассчитаны по следующим формулам:

$$\text{Точность} = \frac{\text{Корректно отвеченные}}{\text{Отвеченные вопросы}};$$

$$\text{Отзывчивость} = \frac{\text{Корректно отвеченные}}{\text{Всего вопросов}}.$$

Результаты эксперимента показывают, что у семантического подхода наряду с подходом описания синтаксических связей как семантических отношений наиболее высокая точность ответов одновременно с низким значением параметра отзывчивости. Такой высокой точности ответов для вопросно-ответной системы, базирующейся на семантическом анализаторе, способствует то, что анализатор может формировать ответы на вопросы с неизвестными типами ожидаемых ответов.

В противовес этому семантический подход лучше применим к вопросам, которые содержат предикаты, из этого следует более строгая стратегия извлечения ответа. Таким образом, большое число вопросов не может быть проанализировано ввиду неоднозначности моделей. На основании результатов также можно предположить, что более высокая точность работы семантического анализатора является следствием того, что он фокусируется лишь на более простых вопросах.

Высокий коэффициент параметра отзывчивости семантико-синтаксического подхода имеет место благодаря большому покрытию синтаксическим компонентом. Поэтому использование семантического извлечения свойств как основного подхода и синтаксического как вспомогательного, когда семантическому подходу не удалось дать какой-либо ответ, является хорошим компромиссом.

В работе рассмотрены актуальные проблемы построения вопросно-ответной системы, в частности, проблема ограниченного объема семантически размеченного корпуса русского языка. Размеченный корпус играет важную роль в работе семантического анализатора, являющегося центральным элементом системы. Точность разбора реализованного семантического анализатора составляет более 80 %. Расширение объемов размеченных корпусов русского языка, по которым происходит обучение анализаторов, должно способствовать повышению качества семантического разбора и как следствие – повышению качества работы вопросно-ответной системы.

Созданная реализация QA-системы формирует ответы в пределах контекста одного предложения, что влияет на точность ее работы. В дальнейшем планируется расширить формирование ответа до контекста абзаца, а затем и всего текста. Для решения данной задачи потребуется расширение базы используемых на данный момент семантических ролей русского языка.

Разработанная система и ее дальнейшее развитие – начальный этап в формировании современных интеллектуальных систем. В настоящее время ведется большое количество разработок, позволяющих анализировать тексты на естественном языке. В частности, обеспечивать отсутствие плагиата в диссертационных и выпускных квалификационных работах, выполнять полнотекстовый поиск по документам и БД, формировать рефераты текстов на естественном языке. На пике развития находятся технологии интеллектуальных помощников, позволяющих анализировать вопросы пользователя и находить ответы с учетом контекста данных вопросов. Следующий шаг пользователя такой системы – начать задавать вопросы

по новостным лентам, статьям, документам и книгам, предполагая получить не целиком статью, документ или книгу, а осмысленный ответ, требующий сопоставления фактов в тексте и формирования выводов.

Литература

1. Запросы как вопросы. URL: https://yandex.ru/company/researches/2012/ya_questions_regions (дата обращения: 12.08.2016).
2. Penas A., Hovy E., Forner P. et al. Question Answering for Machine Reading Evaluation. URL: <http://www.clef-initiative.eu/documents/71612/c076dd78-e36b-40d9-a6c8-fed4b7aa0b3d> (дата обращения: 10.07.2016).
3. START, the world's first Web-based question answering system. Natural Language Question Answering System. URL: <http://start.csail.mit.edu/> (дата обращения: 18.08.2016).
4. Evi. Evi an Amazon Company. URL: <https://www.evi.com/> (дата обращения: 20.08.2016).
5. AQUA. Aktors. Advanced Knowledge Technologies for Shoes. URL: <http://aktors.org/technologies/aqua/> (дата обращения: 4.06.2016).
6. Answers.com – the most trusted place for answering life's questions. Answers Corporation. URL: <http://www.answers.com/> (дата обращения: 14.03.2016).
7. WolframAlpha. Wolfram Alpha LCC. URL: <https://www.wolframalpha.com/> (дата обращения: 15.03.2016).
8. Eхactus Интеллектуальный метапоиск в Интернете. ИСА РАН. URL: <http://exactus.ru/> (дата обращения: 21.08.2016).
9. Смирнов И.В., Шелманов А.О., Кузнецова Е.С., Храмоин И.В. Семантико-синтаксический анализ естественных языков. Ч. II: Метод семантико-синтаксического анализа текстов. М.: Изд-во ИПИ РАН, 2014. URL: http://www.aidt.ru/images/documents/2014-01/11_24.pdf (дата обращения: 28.06.2016).
10. Золотова Г.А., Онипенко Н.К., Сидорова М.Д. Коммуникативная грамматика русского языка. М.: Изд-во РАН, 2004. 544 с.
11. MaltParser 1.9.0 Introduction. MaltParser. URL: <http://www.maltparser.org/intro.html> (дата обращения: 8.06.2016).
12. Graesser A.C., Person N.K. Question asking during tutoring. Memphis State Univ. Pres., 1994, pp. 104–137.
13. Тестовые коллекции РОМИП. Российский семинар по оценке методов информационного поиска. URL: <http://romip.ru/ru/collections/index.html> (дата обращения: 27.08.2016).