ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ
СПбГУТ))) Март 2018 Том 6 № 1 ISSN 2307-1303 http://www.itt.sut.ru
TELECOM IT
March 2018 Vol. 6 Iss. 1 ISSN 2307-1303 http://www.itt.sut.ru

АНАЛИЗ ПОДХОДОВ К СОЗДАНИЮ БАЗЫ ДАННЫХ ВОПРОСНО-ОТВЕТНЫХ СИСТЕМ НА ОСНОВЕ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЕСТЕСТВЕННО ЯЗЫКОВЫХ ТЕКСТОВ

К. В. Ненаусников[1*], С. В. Кулешов[1], А. А. Зайцева[1]

[1] СПИИРАН, Санкт-Петербург, 199178, Российская Федерация
* Адрес для переписки: konstantin2113@mail.ru

Аннотация
Предмет исследования. Статья посвящена проблемам построения вопросно-ответных систем (QA-систем). Предметом исследования являются подходы к автоматическому наполнению базы данных вопросно-ответной системы путем анализа неструктурированных текстовых источников, имеющихся в настоящий момент времени в открытом доступе в сети Интернет. В результате анализа выявлено, что выделяют следующие способы реализации QA-систем: на основе логического вывода по онтологиям, правилам и на основе синтаксиса, с использованием искусственных нейронных сетей. **Метод.** В исследовании разработаны и протестированы методы автоматического выделения вопросно-ответных пар на основе структуры предложений и на основе ассоциативно-онтологического анализа. **Основной результат.** Метод на основе анализа структуры предложений эффективен для текстов типа списков часто задаваемых вопросов (FAQ), а также художественных текстов, содержащих диалоги, прямую речь, основан на предварительной обработке текста, выраженный в виде эвристического правила. Метод на основе ассоциативно-онтологического анализа ориентирован на класс справочных и словарных текстов и основан на предположении о том, что в тексте описательного характера имеется предложение (или группа предложений), содержащее основную мысль текста. В этом случае заголовок текста может считаться вопросом, а это предложение (или группа предложений) – ответом. Для автоматизации выделения смыслообразующих предложений за счет семантической редукции текста применяются алгоритмы реферирования на основе ассоциативно-онтологического подхода к обработке текстов на естественном языке. **Практическая значимость.** Для экспериментальной проверки возможности создания открытой вопросно-ответной системы на базе автоматического сбора вопросно-ответных пар из сети Интернет был разработан прототип модуля сбора базы данных вопросно-ответной системы.

Ключевые слова
Вопросно-ответная пара, ассоциативно-онтологический анализ, текст, автоматическая обработка текста, естественный язык.

Информация о статье
УДК 004.89
Язык статьи – русский.
Поступила в редакцию 21.02.18, принята к печати 28.02.18.

Ссылка для цитирования: Ненаусников К. В., Кулешов С. В., Зайцева А. А. Анализ подходов к созданию базы данных вопросно-ответных систем на основе автоматической обработки естественно языковых текстов // Информационные технологии и телекоммуникации. 2018. Том 6. № 1. С. 92–100.

ANALYSIS OF APPROACHES TO CREATION OF THE BASIS DATA QUESTION-ANSWER SYSTEMS BASED ON AUTOMATIC TREATMENT NATURAL LANGUAGE TEXTS

K. Nenausnikov[1*], S. Kuleshov[1], A. Zaytseva[1]

[1] SPIIRAS, St. Petersburg, 199178, Russian Federation
* Corresponding author: konstantin2113@mail.ru

Abstract—Research subject. The paper is devoted to the problems of question-answer systems constructing (QA-systems). The matter of the study is discussion of approaches to the automatic filling of the database of the QA-system based on the analysis of the unstructured text sources currently available in the public domain of Internet. The analysis revealed that the following ways of implementing QA-systems are distinguished: based on inference for ontologies, rules and syntax, using artificial neural networks. **Method.** The methods for automatically search of question-answer pairs based on the structure of sentences and on the basis of associative-ontological analysis has developed and tested in the research. **Core results.** The method based on the analysis of the structure of sentences is effective for texts such as lists of frequently asked questions (FAQ), as well as literature texts containing dialogs, direct speech, based on preliminary processing of the text, expressed in the form of a heuristic rule. The method based on associative-ontological analysis is focused to the class of reference and dictionary texts and is based on the assumption that in the descriptive text there is a sentence (or a group of sentences) containing the main idea of the text. In this case, the title of the text can be considered a question, and this sentence (or a group of sentences) is the answer. We need to make the selection of meaning-generating sentences due to the semantic reduction of the text automatic. For this purpose, algorithms of self-referencing are applied based on the associative-ontological approach to the processing of texts in natural language. **Practical relevance.** For the experimental verification of the possibility of creating an open QA-system based on the automatic collection of question-answer pairs from the Internet, a prototype of a collection module for the database of the QA-system was developed.

Keywords—question-answer pair, associative-ontological analysis, text, automatic text processing, natural language.

Article info
Article in Russian.
Received 21.02.18, accepted 28.02.18.

For citation: Nenausnikov K., Kuleshov S., Zaytseva A.: Analysis of approaches to creation of the basis data question-answer systems based on automatic treatment natural language texts // Telecom IT. 2018. Vol. 6. Iss. 1. pp. 92–100 (in Russian).

Введение

Классический поиск в сети Интернет позволял эффективно получать интересующие сведения только на ранних этапах развития сети. В настоящее время количество информационных ресурсов и их низкое качество не позволяет использовать только поиск по ключевым словам для ответов на вопросы.

Объем информации, выдаваемой по запросам в популярных поисковых системах, требует от пользователя просмотра больших объемов текста, что в большинстве случаев превышает возможности восприятия человеком за ограниченное время. Более удобным для многих людей видом получения требуемых данных из большого объема текстов может являться диалог с вопросно-ответной системой на естественном языке.

Для вопросно-ответной системы (QA-системы) важно, чтобы эквивалентные по смыслу вопросы распознавались как один и тот же вопрос, независимо от используемых слов, стиля, синтаксических взаимосвязей и идиом. Для поиска или генерации ответа на вопрос QA-система должна иметь доступ к некоторой базе знаний, в которой содержится информация, позволяющая сформировать ответ.

Существуют два основных типа QA-систем: узкоспециализированная (с ограниченной тематической областью) и открытая (не ограниченная конкретной предметной областью). Открытые (общие, или open-domain) QA-системы работают с информацией по всем областям знаний, что обеспечивает возможность вести поиск в смежных областях. Открытая вопросно-ответная система обычно работает с несколькими источниками знаний, в которых производит поиск ответов в зависимости от класса заданного вопроса [1, 2].

Можно выделить следующие способы реализации QA-систем: на основе логического вывода по онтологиям [3], правилам и на основе синтаксиса [4], с использованием искусственных нейронных сетей [5]. Также стоит отметить наличие подходов для повышения качества работы QA-систем на основе показателя удовлетворенности пользователя [6].

Выдаваемый системой ответ по возможности должен быть представлен в виде фразы на естественном языке. В некоторых случаях для этого достаточно простого поиска по базе данных экземпляра коммуникативного акта, в котором был когда-либо использован этот вопрос и дан на него ответ (сформирована вопросно-ответная пара).

Существующие технологии наполнения (формирование базы данных) вопросно-ответных систем включают в себя экспертное наполнение [7], использование краудсорсинговых технологий [8], методы процедурной генерации[1], методы автоматического наполнения с использованием существующих антологий (корпусов текстов).

Рост количества общедоступных информационных ресурсов в сети Интернет, позволяющий обеспечить, с одной стороны, полноту терминологического тезауруса в рамках отдельных предметных областей, а с другой стороны, разнообразие тематических областей, стал основанием сделать предположение о возможности автоматического анализа текстов различного содержания с целью обнаружения и выделения коммуникативных актов для их последующего внесения в базу данных QA-системы в форме вопросно-ответных пар.

[1] Официальный сайт Answers 2009. URL: https://www.wolframalpha.com

Подход к созданию базы данных вопросно-ответной системы

Рассмотрим функциональные особенности QA-системы, позволяющей формировать базу данных (БД) вопросно-ответных пар, извлекая знания из общедоступных ресурсов сети Интернет и предоставляя диалоговый вопросно-ответный интерфейс в форме веб-сервиса (структурная схема приведена на рис. 1). Как видно из рисунка система состоит из функционально независимых блоков формирования базы данных и использования этой базы данных для ответов на запросы пользователя.

За наполнение базы данных отвечает совокупность веб-краулера и модуля сбора вопросно-ответных пар, которые занимаются сбором, загрузкой и анализом текстовых документов, а также извлечением из них вопросно-ответных пар.

За анализ поисковых запросов (текстов вопросов) и выбор наиболее релевантного ответа на этот вопрос среди имеющихся вопросно-ответных пар отвечает интерфейсный поисково-диалоговый компонент, представленный на структурной схеме интерфейсным модулем вопросно-ответной системы.

Формулировку окончательного ответа производит модуль генерации ответов (входящий в состав интерфейсного модуля вопросно-ответной системы), так, чтобы результат выглядел синтаксически естественно и представлял собой именно то, что искал пользователь.

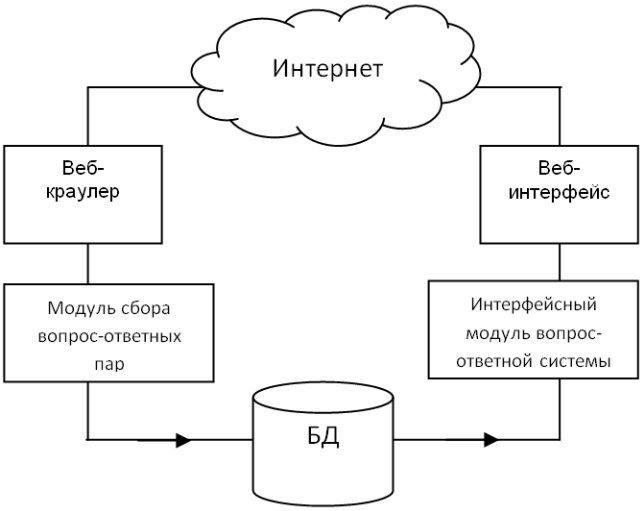

Рис. 1. Структурная схема вопросно-ответной системы

Не углубляясь в механизмы декомпозиции вопроса (пользовательского запроса), поиска и генерации ответа, некоторые из которых рассмотрены, например в [9], рассмотрим только методы автоматического сбора документов для наполнения базы данных (БД) вопросно-ответной системы на основе анализа текстов, доступных в сети Интернет.

Доступные страницы из сети Интернет загружаются с использованием технологии веб-краулера [10], выполняющего обход ссылок в обрабатываемых документах по заданным алгоритмам, в совокупности с headless-браузером, осуществляющим разбор исходного формата загруженного документа (PDF, HTML, MS Word и т. д.) и преобразование его в текстовый формат. Дополнительно из до-

кумента выделяется его заголовок. На данном этапе фильтруются элементы документа, содержащие не относящиеся к основной теме текста блоки информации: рекламные блоки, панели навигации и т. д.

В ходе работы были разработаны и протестированы несколько методов автоматического выделения вопросно-ответных пар на основе структуры предложений и на основе ассоциативно-онтологического подхода к анализу текста [10].

Перед непосредственным выделением вопросно-ответных пар любым из разработанных методов полученные тексты подвергаются предварительной обработке — графематическому анализу [11], включающему в себя определение границ абзацев, предложений и слов, с учетом структуры предложений.

Выделение предложений из текста производится эвристическими правилами на основе поиска символов-разделителей предложений: «.», «!», «?», «…» и символа переноса строки. Границами слов считаются символы-разделители: « », «,», «;», «–», «(», «)», «:» и «”».

При использовании методов на основе ассоциативно-онтологического анализа применяется также предобработка в виде лемматизации — приведения слов к нормальной (словарной) форме и удаления стоп-слов.

Метод выделения вопрос-ответных пар на основе анализа структуры предложений

Для текстов типа списков часто задаваемых вопросов (FAQ), а также художественных текстов, содержащих диалоги, прямую речь, достаточно эффективен метод на основе анализа структуры предложений, полученной при предварительной обработке текста, выраженный в виде следующего эвристического правила. Предложением, содержащим прямую речь, считается предложение, удовлетворяющее любому из условий:

- первый символ предложения – символ «–»;
- внутри предложения последовательно расположена пара символов: первый символ – элемент множества {«,», «.», «!», «?», «”»}, второй – символ «–»;
- внутри предложения последовательно расположена пара символов «:» и «”».

Из полученных предложений удаляются слова автора. Словами автора считается фрагмент текста, который удовлетворяет любому из условий:

- фрагмент текста, стоящий после пары символов: первый символ – элемент множества {«,», «.», «!», «?», «”»}, второй – символ «–»;
- фрагмент текста обособлен символами «–»;
- фрагмент текста, стоящий перед последовательностью символов: «:» и «”».

Предложения, не содержащие прямую речь, не требуют подготовки и рассматриваются в их исходном виде.

Из текста выделяются вопросительные предложения, т. е. предложения, соответствующие следующему условию:

(предложение содержит больше двух слов) И (предложение оканчивается символом «?»).

Непосредственно после вопросительных предложений в пределах одного абзаца выделяется предложение, удовлетворяющие условиям:

(предложение не должно оканчиваться символом «?») И (предложение содержит не менее одного слова).

Такое предложение будем считать ответом на поставленный вопрос. Если ни одно предложение в данном абзаце не удовлетворяет этим условиям, то считаем, что вопрос не содержит ответа и не заносится в базу данных.

Данные эвристические правила могут быть записаны в виде порождающей грамматики и реализованы в виде конечного автомата.

Метод выделения вопросно-ответных пар на основе ассоциативно-онтологического анализа

Метод на основе ассоциативно-онтологического анализа в первую очередь ориентирован на класс справочных и словарных текстов и основан на предположении о том, что в тексте описательного характера имеется предложение (или группа предложений), содержащее основную мысль текста. В этом случае заголовок текста (в том числе, обозначенный через мета-теги интернет-документа) может считаться вопросом, а это предложение (или группа предложений) – ответом.

Использование алгоритмов реферирования на основе ассоциативно-онтологического подхода к обработке текстов на естественном языке [12] позволяет автоматизировать выделение смыслообразующих предложений за счет семантической редукции текста. Реферирование текстов выделяет группу предложений на основе биграмм слов, где под биграммой понимается пара слов, встречающихся в одном предложении. Пару слов, достаточно часто находящихся в одном предложении, считают ассоциативно связанной, причем чем чаще встречается эта биграмма, тем сильнее связь. Предложения, содержащие понятия, сумма связей которых наибольшая, лучше всех предложений отражают предметную область, описанную в тексте.

Экспериментальная проверка и оценка результатов

Для экспериментальной проверки возможности создания открытой вопросно-ответной системы на базе автоматического сбора вопросно-ответных пар из сети Интернет был разработан прототип модуля сбора, работающий совместно с веб-краулером системы мониторинга интернет-ресурсов [12]. Системой было обработано 310 239 документов с полезным (без учета разметки документа и медиаданных) объемом текста 1,92 Гбайт. При анализе текстов получено 2 230 325 вопросов и ответов к ним, объем базы данных составил 710 Мбайт. Количественные результаты, полученные при экспериментальной проверке различных методов, представлены в табл. (см. ниже).

Наибольший вклад в формирование БД вопросно-ответных пар среди текстов, содержащих зафиксированные коммуникативные акты, внесли, в основном благодаря высокому удельному содержанию вопросно-ответных пар внутри каждого документа:

- художественные тексты, содержащие диалоги героев (26 %);
- разделы часто-задаваемых вопросов (FAQ) (17 %);
- справочные и словарные источники при использовании алгоритма на основе ассоциативно-семантического анализа (21 %);

- контент, генерируемый пользователем (UGC – *user generated content*): форумы, блоги, комментарии;
- тексты документального и новостного характера.

Таблица.

Количество полученных вопросно-ответных пар

Метод	**Количество выделенных вопросно-ответных пар**
Метод на основе анализа структуры предложения без учета прямой речи	529 117
Метод на основе анализа структуры предложения для прямой речи	1 080 730
Метод на основе ассоциативно-онтологического анализа	310 239

Заключение

В процессе исследования был разработан прототип системы сбора вопросно-ответных пар на основе фактического материала, содержащегося в открытом доступе в сети Интернет.

Доступные страницы загружались с использованием технологии веб-краулера, выполняющего обход ссылок в совокупности с headless-браузером, осуществляющим разбор исходного формата загруженного документа.

Были проверены два метода для выделения вопросно-ответных пар: метод, основанный на анализе структуры предложений, и метод на основе ассоциативно-онтологического анализа текстов.

На основе анализа результатов, полученных разработанными методами, можно утверждать, что для конкретной выборки среднее количество вопросно-ответных пар составило 7,9 на 1 документ (одна вопросно-ответная пара на 1 килобайт текста).

В то же время, экспертная оценка качества и полноты базы данных, проведенная с использованием диалогового прототипа, показала невозможность получения адекватных ответов по большинству заданных поисковых запросов, которые эксперт задавал поисковой системе без учета предметной области.

Это говорит об ограниченности возможностей по созданию открытой (не узкоспециализированной) QA-системы только путем непосредственного выделения вопросно-ответных пар из неструктурированных текстовых источников, имеющихся в настоящий момент времени в открытом доступе в сети Интернет.

Исследование проводится при частичной поддержке гранта РФФИ №16-29-12965.

Литература

1. Лапшин В. А. Вопросно-ответные системы: развитие и перспективы // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2012. № 6. С. 1–9.

2. Rodrigo A., Peñas A. A Study about the Future Evaluation of Question-Answering Systems // Knowledge-Based Systems. 2017. Vol. 137. pp. 83–93.

3. Zou L., Huang R., Wang H., Yu J. X., He W., Zhao D. Natural Language Question Answering over RDF: A Graph Data Driven Approach // International Conference on Management of Data (SIGMOD). 2014. pp. 313–324.

4. Fader A., Zettlemoyer L., Etzioni O. Open Question Answering over Curated and Extracted Knowledge Bases // International Conference on Knowledge Discovery and Data Mining, (SIGKDD). 2014. pp. 1156–1165.

5. Li J., Liu H., Zhang Y., Xing C. A Health QA with Enhanced User Interfaces // 13th Web Information Systems and Applications Conference (WISA). 2016. pp. 173–178.

6. Liu Y., Bian J., Agichtein E. Predicting Information Seeker Satisfaction in Community Question Answering // 31st Annual International Conference on Research and Development in Information Retrieval (SIGIR). 2008. pp. 483–490.

7. Сутягин И. В. Методы формализации экспертных знаний для наполнения базы знаний // Молодой ученый. 2012. Т. 1. № 1 (36). С. 151–153.

8. Федоркова Г. С. Краудсорсинговые технологии в российских социальных медиа // Всероссийская научно-практическая конференция «Коммуникации в современном мире». 2017. С. 154–155.

9. Никитин А., Райков П. Вопросно-ответные системы. URL: http://yury.name/internet/06ia-seminar.ppt

10. Кулешов С. В, Зайцева А. А., Марков В. С. Ассоциативно-онтологический подход к обработке текстов на естественном языке // Интеллектуальные технологии на транспорте. 2015. № 4. С. 40–45.

11. Первушин А. Модуль графематического анализа в системе обработки русскоязычных текстов // Новые информационные технологии в автоматизированных системах. 2012. № 15. С. 187–190.

12. Александров В. В., Кулешов С. В. Аналитический мониторинг Internet-контента. Инфологический подход // Качество. Инновации. Образование. 2008. № 3 (34). С. 68–70.

References

1. Lapshin V. QA Systems: Development and Prospects // Nauchno-tekhnicheskaya informatsiya. Seriya 2. Informatsionnye protsessy i sistemy. 2012. Iss. 6. pp. 1–9.

2. Rodrigo A., Peñas A. A Study about the Future Evaluation of Question-Answering Systems // Knowledge-Based Systems. 2017. Vol. 137. pp. 83–93.

3. Zou L., Huang R., Wang H., Yu J. X., He W., Zhao D. Natural Language Question Answering over RDF: A Graph Data Driven Approach // International Conference on Management of Data (SIGMOD). 2014. pp. 313–324.

4. Fader A., Zettlemoyer L., Etzioni O. Open Question Answering over Curated and Extracted Knowledge Bases // International Conference on Knowledge Discovery and Data Mining, (SIGKDD). 2014. pp. 1156–1165.

5. Li J., Liu H., Zhang Y., Xing C. A Health QA with Enhanced User Interfaces // 13th Web Information Systems and Applications Conference (WISA). 2016. pp. 173–178.

6. Liu Y., Bian J., Agichtein E. Predicting Information Seeker Satisfaction in Community Question Answering // 31st Annual International Conference on Research and Development in Information Retrieval (SIGIR). 2008. pp. 483–490.

7. Sutyagin I. Methods for Formalizing Expert Knowledge for Filling the Knowledge Base // Molodoi uchenyi. 2012. Vol. 1. Iss. 1 (36). pp. 151–153.

8. Fedorkova G. Crowdsourcing Technologies in Russian Social Media // All-Russian Scientific-Practical Conference «Communications in the Modern World». 2017. pp. 154–155.

9. Nikitin A., Raikov P. QA Systems. URL: http://yury.name/internet/06ia-seminar.ppt

10. Kuleshov S, Zaytseva A., Markov V. Associative-Ontological Approach to Natural Language Texts Processing // Intellektual'nye tekhnologii na transporte. 2015. Iss. 4. pp. 40–45.

11. Pervushin A. Module of Graphematic Analysis in the System for Processing Russian-Language Texts // Novye informatsionnye tekhnologii v avtomatizirovannykh sistemah. 2012. Iss. 15. pp. 187–190.

12. Aleksandrov V., Kuleshov S. Analytical Monitoring of Internet Content. Infological Approach// Kachestvo. Innovatsii. Obrazovanie. 2008. Iss. 3 (34). pp. 68–70.

Ненаусников Константин Вячеславович – аспирант, СПИИРАН, Санкт-Петербург, 199178, Российская Федерация, konstantin2113@mail.ru

Кулешов Сергей Викторович – доктор технических наук, СПИИРАН, Санкт-Петербург, 199178, Российская Федерация, kuleshov@iias.spb.su

Зайцева Александра Алексеевна – кандидат технических наук, старший научный сотрудник, СПИИРАН, Санкт-Петербург, 199178, Российская Федерация, cher@iias.spb.su

Nenausnikov Konstantin – Postgraduate, SPIIRAS, St. Petersburg, 199178, Russian Federation, konstantin2113@mail.ru

Kuleshov Sergey – Doctor of Engineering Sciences, Full Professor, SPIIRAS, St. Petersburg, 199178, Russian Federation, kuleshov@iias.spb.su

Zaytseva Alexandra – Candidate of Engineering Sciences, Senior Research Officer, SPIIRAS, St. Petersburg, 199178, Russian Federation, cher@iias.spb.su