РОССИЙСКАЯ АКАДЕМИЯ НАУК

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
(ВИНИТИ)

НАУЧНО • ТЕХНИЧЕСКАЯ
ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 10 Москва 2013

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

УДК 004.5

А. А. Соловьёв, О. В. Пескова

Об архитектурах программных систем вопросно-ответного поиска

Представлен анализ принципов построения известных систем вопросно-ответного поиска, позволяющего пользователю получать лаконичный ответ на вопрос, сформулированный на естественном языке. Предложена классификация архитектур вопросно-ответных систем, рассмотрены их основные характеристики, выделены достоинства и недостатки.

Ключевые слова: *вопросно-ответный поиск, классификация архитектур вопросно-ответных систем, информационный поиск, экспертные системы, программное обеспечение*

ВВЕДЕНИЕ

Программные системы вопросно-ответного поиска, или просто вопросно-ответные системы (от англ: *Question Answering Systems*), – это вид информационно-поисковых систем, способных обрабатывать введенный пользователем вопрос на естественном языке и выдавать осмысленный ответ. В отличие от задачи классического поиска по ключевым словам, в которой результатом является перечень документов, содержащих ответ на вопрос, в задаче вопросно-ответного поиска результат – это краткий и лаконичный ответ, сформированный системой в результате анализа разнообразных источников данных. Примером такого источника может служить некоторая коллекция полнотекстовых документов (множество страниц глобальной сети Интернет), а ответ составляется из фрагмента наиболее релевантного документа коллекции.

Традиционно в работах по вопросно-ответному поиску приводят классификацию методов или систем по используемому математическому аппарату:

- логические формы и логический вывод,
- графы зависимостей слов в предложениях,
- статистический подход и машинное обучение классификаторов,
- лексические онтологии для анализа отдельных слов текста и др.

Этой классификации следует в своём обзоре 2006 г. J. Prager [1]. Более поздний обзор других авторов [2] также следует этой классификации. Боль-

шинство исследователей следуют некоторой типовой архитектуре вопросно-ответной системы, которая, по сути, заключается в разбиении задачи вопросно-ответного поиска на четыре подзадачи: анализ вопроса, поиск, извлечение потенциальных ответов и валидация ответов. С точки зрения проектирования программного комплекса архитектурой называют способ разбиения системы на модули и определение связей между ними. В настоящей работе предлагается классификация архитектур вопросно-ответных систем, т.е. способов разбиения систем на модули.

Существует ряд различных подходов и принципов построения вопросно-ответных систем (ВОС), основными из которых являются следующие:

1) метапоисковая система;
2) система поиска по аннотированному тексту;
3) экспертная система;
4) система поиска в коллекциях вопросов и ответов.

Далее приводится обзор перечисленных архитектур (как способов декомпозиции ВОС на составляющие компоненты и подзадачи) и примеры известных программных реализаций рассмотренных подходов.

МЕТАПОИСКОВАЯ СИСТЕМА

Архитектура метапоисковой системы предусматривает использование существующей классической поисковой системы в качестве источника данных (рис. 1). ВОС преобразует введённый пользователем вопрос на естественном языке в запрос в виде ключевых слов и анализирует выдачу поисковой системы – несколько наиболее релевантных документов или их фрагментов (сниппетов).

Система анализирует вопрос пользователя с целью выделить следующие данные [3]:

- предположение о *семантическом классе ответа*;
- *фокус вопроса*, т. е. вопросительные слова, обозначающие искомую информацию, например, «в каком городе», «кто», «где», «в каком году», «когда», «сколько», «сколько метров», «какой высоты», «какого цвета» и др.; в случае простого фактографического вопроса искомая информация обычно является каким-то атрибутом некоторого объекта (имя, вес, цвет);
- *опора вопроса*, т. е. остальные члены вопросительного предложения, которые описывают уникальные свойства искомого объекта.

Метапоисковые ВОС обычно формулируют запрос по ключевым словам на основе слов, входящих в *опору вопроса*. В англоязычных системах распространён приём расширения поискового запроса синонимами и гипонимами на основе лексической онтологии WordNet [4].

Результаты поиска по ключевым словам – сниппеты – обрабатываются существующими компонентами систем автоматической обработки текста (компьютерной лингвистики). Например, выделяются все именованные сущности, соответствующие искомому семантическому классу: персоны, топонимы (географические названия), названия организаций, линейные размеры и т.п. Более глубокий лингвистический анализ (например, синтаксический или семантический разбор) позволяет выбрать из всех найденных сущностей нам более подходящие [5].

Преимущества метапоисковой архитектуры заключаются в следующем:

- отсутствие собственного индекса документов и, как следствие, отсутствие необходимости хранить огромный массив информации (для поиска в Интернете);

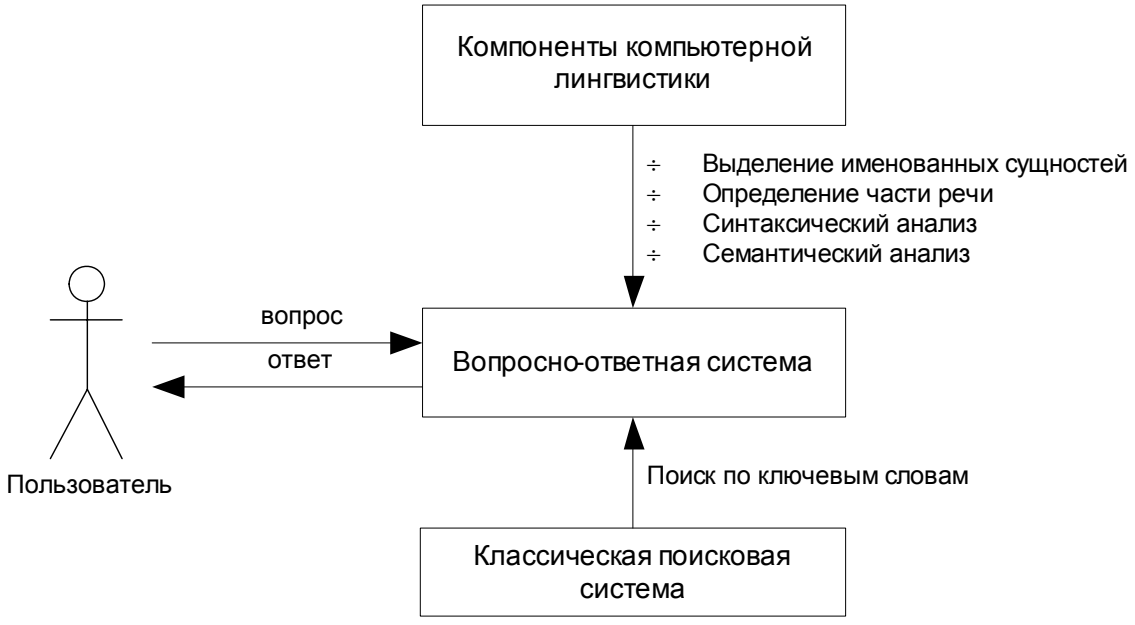

Рис. 1. Обобщённая схема архитектуры метапоисковой вопросно-ответной системы

- гибкость, т.е. возможность абстрагироваться от задач поиска и компьютерной лингвистики. ВОС использует поисковую машину и лингвистические компоненты как чёрные ящики, и обычно не возникает проблем с заменой этих компонентов. Метапоисковая ВОС может использовать любые доступные инструменты для анализа сниппетов, независимо от математического аппарата, используемого в поисковой машине или лингвистических компонентах. Например, поисковая машина может работать на деревьях решений, построенных методами машинного обучения, синтаксический анализ может выполняться вероятностными методами, а ВОС в то же время будет анализировать вопрос с помощью регулярных выражений и представлять сниппеты в виде графов синтаксических зависимостей.

Недостатками метапоисковой архитектуры являются:

- высокая вычислительная нагрузка в момент обработки вопроса, введённого пользователем, связанная с высокими вычислительными затратами на выполнение лингвистических задач;
- ограничения по управлению поиском (длина и «целостность» сниппетов, авторитетность источников и др.).

ПОИСК ПО АННОТИРОВАННОМУ ТЕКСТУ

Вопросно-ответные системы, построенные по принципу поисковых систем с коллекциями аннотированных документов, имеют в своём составе поисковый индекс документов (рис. 2). Данный индекс, в отличие от классических поисковых систем, дополняется специфическими для ВОС атрибутами. Элементами индекса являются не отдельные слова текста, а объекты детального лингвистического анализа, например:

- именованные сущности [5];
- элементарные синтаксические связки (пары грамматически связанных слов и др.);
- предикативно-аргументные структуры [6].

Построение индекса происходит с привлечением компьютерной лингвистики: каждый новый документ проходит автоматическую обработку текста на естественном языке, размечаются требуемые ВОС объекты, затем они добавляются в индекс.

Использование своего специального индекса позволяет преодолеть некоторые недостатки метапоисковой архитектуры.

Преимуществами поиска по аннотированному тексту являются:

- меньшая (по сравнению с метапоисковыми ВОС) вычислительная нагрузка в момент обработки вопроса пользователя в реальном времени благодаря специализированному индексу;
- возможность, благодаря специализированному индексу, организовать наиболее удобный для ВОС поисковый аппарат.

Недостатки поиска по аннотированному тексту состоят в следующем:

- невысокая гибкость по сравнению с метапоисковой системой: на этапе построения индекса выбирается какая-то определённая модель представления текста. Любые изменения, скорее всего, потребуют перестройки индекса – например, изменение онтологии семантических классов и имён объектов или замена грамматики зависимостей (*dependency grammar)* грамматикой связей*,(link grammar*) [7] в модели представления текста;

Рис. 2. Обобщённая схема архитектуры вопросно-ответной системы, основанной на поиске по аннотированным текстам

- потребность в значительно бóльших вычислительных ресурсах, связанная с необходимостью индексации всей коллекции (по сравнению с классическим поиском). При этом данные ресурсы расходуются недостаточно эффективно: документы анализируются целиком, хотя ответы на вопросы пользователей содержатся в очень редких предложениях.

ЭКСПЕРТНАЯ СИСТЕМА

Вопросно-ответные системы, построенные по принципу работы со структурированными базами данных, можно отнести к классу экспертных систем (рис. 3). Вопрос на естественном языке преобразуется в поисковый запрос к базе данных, содержащей структурированные факты, например, фреймы. В отличие от рассмотренных выше архитектур, такие системы могут выполнять логический вывод новой информации на основе множества разрозненных фактов. Метапоисковые же системы и системы со специальным индексом – это полнотекстовые поисковые системы. Они ищут ответ на вопрос, очевидным образом содержащийся в полном тексте документа, не пытаясь делать логический вывод новой информации, не содержащейся в тексте в явном виде.

База данных фактов может быть построена автоматически в результате анализа коллекции документов. Этот процесс аналогичен построению аннотированного индекса. Однако он происходит на более детальном уровне обработки естественного текста: извлекаются не синтаксические или поверхностно-семантические конструкции, а факты. Такие системы могут не хранить текст исходного документа, из которого был извлечён той или иной факт (фрейм).

Преимуществами ВОС, спроектированных на основе экспертной системы, являются:

- высокая скорость работы (по сравнению, например, с метапоисковой архитектурой);
- точность и достоверность результатов.

Недостатки заключаются в следующем:

- сильная зависимость от структуры фактов (фреймовой модели). Одна структура может быть адекватна одной предметной области, но не другой. Обычно это требует постоянной работы аналитиков-редакторов, чёткого понимания потребностей пользователей и очень внимательной проработки модели данных, постоянного расширения и модификации это модели для новых предметных областей;
- необходимость выбирать только авторитетные исходные тексты для извлечения информации об окружающем мире. При этом извлечённые факты могут противоречить друг другу и система должна это учитывать;
- трудоёмкость построения базы фактов, причём как вычислительная, так и организационная. Лингвистическая обработка на глубоком семантическом уровне (например, на уровне извлечения фактов) подвержена большому количеству ошибок. Она аккумулирует все ошибки лингвистической обработки на предшествующих уровнях: графематическом, морфологическом, синтаксическом и поверхностно-семантическом.

Рис. 3. Обобщённая схема архитектуры вопросно-ответной системы, применяющей принципы построения экспертных систем

ПОИСК В КОЛЛЕКЦИИ ВОПРОСОВ И ОТВЕТОВ

Другим подходом к автоматизации поиска ответов на вопросы пользователей является социальный вопросно-ответный поиск (*collaborative question answering*). В таких системах одни пользователи отвечают на вопросы других. Пользователь, имеющий информационную потребность, открывает страницу веб-сайта системы и формулирует вопрос. Система ищет похожие вопросы в коллекции вопросов и ответов и выдаёт найденный раздел, где обсуждается вопрос. Если подобного вопроса не существует, создаётся новый раздел для обсуждения вопроса. На этот вопрос отвечают другие пользователи, автору вопроса приходят уведомления по мере появления ответов. Данные в такой системе представлены в виде коллекции вопросов с ответами, которая может пополняться другими пользователями или даже автоматически.

Усложнением системы является модуль извлечения вопросов и ответов из коллекции документов. ВОС непрерывно сканирует все страницы Интернета, анализирует тексты на естественном языке и формулирует возможные вопросы по этому тексту (см. *LCC Predictive questioning* в [5]). Аналогично социальным ВОС, где пользователи оценивают ответы друг друга, данная модификация позволяет поднимать или понижать рейтинг автоматически сгенерированной пары «вопрос–ответ».

Другой подход предлагают разработчики немецкой системы LogAnswer [8]: система работает как программный робот, который обходит известные вопросно-ответные сайты с пользовательским содержимым и пытается отвечать на вопросы автоматически как обычный участник обсуждений.

Преимуществами использования коллекций вопросов и ответов являются:

- возможность развёрнутых, необязательно фактографических ответов;
- проверка достоверности ответов другими пользователями;
- низкие вычислительные затраты на поиск ответа в коллекции.

Недостатками являются:

- необходимость мотивации пользователей как для пополнения коллекции, так и для простановки оценок, особенно для ответов и вопросов, порождённых автоматически;
- трудоёмкость автоматического порождения коллекции, необходимость объёмного хранилища.

ОБЗОР ИЗВЕСТНЫХ ВОПРОСНО-ОТВЕТНЫХ СИСТЕМ

В таблице приведены примеры исследовательских и коммерческих систем, классифицированных по рассмотренным архитектурам.

Рис. 4. Обобщённая схема архитектуры системы социального вопросно-ответного поиска

Таблица

Известные примеры исследовательских и коммерческих ВОС

		Исследовательские	Коммерческие
Метапоиск		LCC Power Answerer [9], SMU Falcon, OpenEphyra [6], AskMsr, AnswerBus, Умба [10]	–
Аннотированный индекс		LCC Chaucer-2 [101], IBM Watson [12], Javelin, START [13]	Powerset, Asknet [14]
Структурированная БД		BASEBALL, LUNAR	WolframAlfa [15], TrueKnowledge
Коллекция вопросов и ответов	**Пополняется пользователями**	–	Chacha, Ask.com [16],Yahoo!Answers [17], otvety.google.com, ответы@mail.ru
	Пополняется автоматически	LCC Ferret [5]	Swingly

Отметим отсутствие коммерческих реализаций метапоисковых систем. Это объясняется высокими вычислительными затратами на обработку каждого запроса. Например, системам OpenEphyra и Умба может потребоваться до нескольких минут процессорного времени на обработку одного вопроса. Отметим также, что система IBM Watson, успешно выступающая в телевикторине Jeopardy, работает на кластере из 3000 узлов. Создатели отмечают, что на одном процессоре системе требуется несколько часов на ответ. Однако AskNet, использующая собственный индекс, выдаёт ответ за доли секунды.

Особый интерес представляет семейство систем от Language Computer Corporation (LCC). Компания экспериментировала со всеми архитектурами и недавно запустила коммерческий стартап Swingly, основанный на исследовательских прототипах компании.

ЗАКЛЮЧЕНИЕ

В результате проделанной работы выполнен анализ подходов к решению задачи автоматического поиска ответа на вопрос, сформулированный пользователем на естественном языке. Выделены четыре типа архитектуры ВОС, у каждой из которых есть своё соотношение между трудоёмкостью предобработки информации и вычислениями «на лету», между более низким, но надёжным уровнем лингвистической обработки языка и более высоким уровнем абстракции. Некоторые архитектуры ориентированы на поиск ответа, который присутствует в явном виде, другие же позволяют порождать новую информацию на основе логического вывода из доступных фактов. Поиск ответа в полном тексте остаётся вычислительно сложной задачей: некоторым исследовательским системам требуется от нескольких минут до нескольких часов работы одного процессора на поиск ответа на вопрос. Например, IBM Watson потребовалось 3000 процессоров, чтобы успешно конкурировать с людьми в телевикторине. В то же время – это пример наиболее гибкой вопросно-ответной системы, которая использует полный арсенал современных методов для решения задачи.

СПИСОК ЛИТЕРАТУРЫ

1. Prager John. Open-Domain Question–Answering // Foundation and Trends in Information Retrieval. – 2006. – Vol. 1, № 2. – P. 91–231.
2. Kolomiyets Oleksandr, Moens Marie-Francine. A survey on question answering technology from an information retrieval perspective // Information Sciences. – 2011. – Vol. 181, Is. 24. – P. 5412–5434.
3. Соловьёв А. А., Пескова О. В. Построение вопросно-ответной системы для русского языка: модуль анализа вопросов // Новые информационные технологии в автоматизированных системах: материалы 13-го научно–практического семинара. – Моск. гос. ин-т электроники и математики. – 2010. – С. 41–49. – URL: http://nps.itas.miem.edu.ru/2010/sbornik13.pdf, свободный.
4. Проект WordNet. – URL: http://wordnet.princeton.edu/, свободный.
5. Harabagiu S., Hickl A., Lehmann J., Moldovan D. Experiments with interactive question–answering // Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05). Association for Computational Linguistics. Stroudsburg, PA. – 2005. – P. 205–214.
6. Schlaefer N. A Semantic Approach to Question Answering. – Saarbrücken, 2007. – P. 27.
7. Протасов С. В. Обучение с нуля грамматике связей русского языка // X Национальная конференция по искусственному интеллекту с международным участием «КИИ-06». – М., 2006. – С. 515–524.
8. Dong T., Furbach U., Glöckner I., Pelzer B. A natural language question answering system as a participant in human Q&A portals // Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI–2011). Barcelona, Spain. July 2011. – P. 2430–2435.
9. Moldovan D., Pasca M., Surdeanu M. Some Advanced Features of LCC's Poweranswer // Advances in Open Domain Question Answering.

Text, Speech and Language Technology. – 2006. – Vol. 32, Part 1. – P. 3–34.

10. Соловьёв А. А. Кто виноват и где собака зарыта? Метод валидации ответов на основе неточного сравнения семантических графов в вопросно-ответной системе // Российский семинар по оценке методов информационного поиска. Труды РОМИП. – Казань, 2010. – С. 125–141.
11. Hickl A., Roberts K., Rink B., Bensley J., Jungen T., Shi Y., Williams J. Question Answering with LCC's Chaucer-2 at TREC 2007 // Proceedings of TREC 2007. Gaithersburg, MD, 2007.
12. Ferrucci D., Brown E., Chu–Carroll J., Fan J., Gondek D., Kalyanpur A. A., Lally A., Murdock J. W., Nyberg E., Prager J., Schlaefer N.,Welty Ch. Building Watson: An overview of the DeepQA project // AI Magazine. – 2010, 31(3). – P. 59–79.
13. Katz B., Borchardt G., Felshin S. Natural Language Annotations for Question Answering // Proceedings of the 19th International FLAIRS Conference (FLAIRS 2006). May 2006. Melbourne Beach, FL, 2006. – P. 303–306.
14. Проект AskNet. – URL: http://asknet.ru/, свободный.
15. Проект WolframAlpha. – URL: http://www.wolframalpha.com/, свободный.
16. Проект Ask.Com. – URL: http://www.ask.com/, свободный.
17. Проект Yahoo!Answers. – URL: http://answers.yahoo.com/, свободный.

Материал поступил в редакцию 20.06.13.

Сведения об авторах

СОЛОВЬЁВ Александр Александрович – аспирант МГТУ им. Н.Э. Баумана, Научно-техническая библиотека, программист
e-mail: a-soloviev@mail.ru

ПЕСКОВА Ольга Вадимовна – кандидат технических наук, доцент МГТУ им. Н.Э.Баумана,
e-mail: opeskova@mail.ru