DOI 10.58351/2949-2041.2025.24.7.011

Рудаков Игорь Владимирович, к.т.н., доцент
МГТУ им. Н. Э. Баумана

Волкова Лилия Леонидовна, старший преподаватель
МГТУ им. Н. Э. Баумана

Глотов Илья Анатольевич, магистрант
МГТУ им. Н. Э. Баумана

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СУЩЕСТВУЮЩИХ ПОДХОДОВ К ПОСТРОЕНИЮ ВОПРОСНО-ОТВЕТНЫХ СИСТЕМ

Аннотация. Проведен анализ предметной области автоматической обработки текстов на естественном языке в части этапов анализа и синтеза текстов и выполнен сравнительный анализ существующих подходов к построению вопросно-ответных систем.

Ключевые слова: Вопрос-ответные системы, диалоговые системы, графематический анализ, морфологический анализ, синтаксический анализ, семантический анализ, метапоисковая система, экспертная система.

Введение. Одним из важнейших направлений автоматической обработки естественно-языковых данных является разработка и совершенствование интеллектуальных диалоговых систем и их упрощенных версий – чат-ботов. Эти системы стали все чаще применяться в коммерческих проектах, где они используются в общении с клиентами для помощи в покупке товаров, технической поддержки, навигации по сайтам и т. д [1]. Диалоговые системы используются в качестве интеллектуальных модулей коммуникации социальных роботов [2].

Анализ текстов. Компоненты, составляющие структуру программных реализаций методов анализа текстов, – лингвистические процессоры, которые последовательно обрабатывают входной текст. Вход одного процессора является выходом другого [3]. Последовательность обработки текста связана с уровнями анализа текста.

1. Графематический анализ – выделение словоформ (слов), составных знаков пунктуации, цифровых комплексов, формул и т. д.
2. Морфологический анализ – определение морфологических характеристик словоформ входного текста.
3. Синтаксический анализ – построение дерева синтаксических зависимостей одного предложения.
4. Семантический анализ – построение семантического представления текста.

Для каждого метода анализа текста разрабатывается свой язык представления текста на соответствующем уровне. Язык представления состоит из констант и правил их комбинирования. На графематическом уровне константами являются графематические дескрипторы (ЛЕ – лексема, ЦК – цифровой комплекс и т. д.), на морфологическом уровне – граммемы (рд – родительный падеж, мн – множественное число и т. д.), на синтаксическом – названия отношений (subj – отношение между подлежащим и сказуемым, circ – обстоятельство и т. д.).

Уровням анализа текста сопоставляют модули программного обеспечения по разбору текстов на естественном языке. Эти модули последовательно анализируют данные текста на каждом уровне; таким образом результаты работы предыдущего модуля служат входными данными следующего модуля. Следует отметить, что последующие компоненты также могут улучшить представление предыдущих, поэтому, в частности, модуль синтаксического анализа текста разделяют на три: предсинтаксический, собственно синтаксический и постсинтаксический [4].

Графематический анализ. Метод графематического анализа создает основу для последующего морфологического и синтаксического анализа на основе выделения слов, цифровых комплексов, формул. Анализ направлен на разбивку текста на слова, разделители; сборку слов, написанных вразрядку; выделение устойчивых оборотов, фамилии, имени, отчества, даты; выделение электронных адресов и имен файлов; выделение из входного текста предложений, абзацев, заголовков, примечаний [5].

На вход компоненту графематического анализа подается текст, на выходе строится графематическая таблица, в которой на каждой строке стоит слово или разделитель из входного текста [3].

Графематическая таблица состоит из двух столбцов. В первом столбце стоит некоторый фрагмент входного текста, во втором столбце стоят графематические дескрипторы, характеризующие этот фрагмент текста.

Например, таблица 1 представляет графематическую таблицу для текста «Александр спал».

Таблица 1

Пример графематической таблицы

Фрагмент входного текста	Графематические дескрипторы
Александр	ЛЕ Бб ПРД1
_	РЗД ПРБ
Спал	Ле бб ПРД2

Дескрипторы создают формальное описание текста на уровне графе-матики, которое уже поддаётся автоматизированной обработке в терминах лингвистических теорий.

Морфологический анализ. Метод морфологического анализа осуществляет морфоанализ и лемматизацию словоформ [3]. Морфологический анализ – приписывание словоформам морфологических признаков, лемматизация – приведение текстовых форм слова к словарным. При лемматизации для каждого слова входного текста морфологический процессор выдает множество морфологических интерпретаций – троек следующего вида: лемма; морфологическая часть речи; множество наборов граммем.

Например, словоформе *стол* с леммой СТОЛ будут приписаны следующий набор граммем: «мр, ед, им», «мр, ед, вн». Таким образом, морфологический анализ выдает два варианта анализа словоформы *стол* с леммой СТОЛ внутри одной морфологической интерпретации: с винительным (вн) и именительным падежами (им). Также большую роль здесь играет омонимичность словоформ. Например, у словоформы *стали* могут быть интерпретации: сталь – существительное; стать – глагол.

Таким образом, видно, что морфологического анализа явно не достаточно для выбора одной конкретной морфологической интерпретации слова, к тому же, выбор одной интерпретации может повлиять на выбор интерпретации для соседних слов. Поэтому программы работают с целым набором возможных морфологических интерпретаций, постепенно выделяя наиболее вероятные на следующих этапах анализа [3].

Синтаксический анализ. Цель синтаксического анализа – построение групп на предложении. Синтаксическая группа – это отрезок (первое слово группы – последнее слово группы) в предложении, для которого указан подотрезок – его главная группа. В частном случае группа – одно слово. Как видно из определения, синтаксические группы неразрывны, а из того, что две группы пересекаются, следует, что одна лежит в другой (т. е. является ее подотрезком). Синтаксическую структуру предложения можно представить в виде дерева: корень (нулевой уровень) – само предложение; узлы – синтаксические группы (далее просто группы); листья – элементарные группы (слова); ребра – отношение «лежать непосредственно в» (A ⇒ B значит, что B лежит в A и при этом нет такой группы C, что B лежит в C и C лежит в A). До начала работы анализатора каждое слово – группа первого уровня (группы первого уровня не входят ни в какие группы кроме предложения) и, кроме корня, других групп нет.

Результатом работы является дерево синтаксических зависимостей предложения, описывающее лингвистические отношения подчинения между словами, представленными словоформами и составляющими предложение. По сути, это и есть математическая модель предложения на естественном языке [3].

Синтаксический анализ может быть выполнен при помощи механизма разбора на основе грамматик [6; 7]. Если одна или несколько словоформ омонимичны, то есть имеют несколько вариантов разбора, анализируют различные варианты формирования деревьев синтаксических зависимостей с последующим отбрасыванием «неудачных» деревьев, которые не удалось сформировать как связные графы [8].

Семантический анализ. Семантический (смысловой) анализ необходим для оценивания смысла передаваемой информации, соотношения ее с информацией, которая хранилась до появления обрабатываемой информации. На этом этапе формируется некоторое представление смысла текста на естественном языке. Семантические связи между словами или другими единицами языка отражаются в семантических словарях [9]. Основными задачами семантического анализа являются построение семантической интерпретации слов и лингвистических конструкций (например, предложений) и установление семантических отношений между различными элементами текста.

При семантическом анализе предложений используют падежные грамматики и семантические валентности, а семантика предложения задается через связи главного слова (глагола) с его семантическими актантами [10].

Основой семантического анализа является утверждение, что конкретное значение слова не является элементарной семантической единицей. Оно, в свою очередь, делится на более мелкие единицы – единицы семантического словаря языка, являющиеся своеобразными атомами, комбинации которых складываются в «молекулы» – значения слов естественного языка. Именно семантический анализ дает возможность решить проблемы многозначности (омонимии), которая часто возникает при автоматическом анализе на разных языковых уровнях [9].

Синтез текстов. Синтез текстов на естественном языке можно считать задачей, обратной задаче анализа (синтаксического, семантического) предложения на естественном языке. Целью синтеза является построение предложения на естественном языке по полученному дереву с учетом грамматических и синтаксических правил целевого языка [12]. Синтез текста выполняется в порядке, обратном последовательности этапов анализа текстов: семантический синтез, синтаксический синтез, морфологический синтез, графематический синтез [4].

Обзор подходов к построению вопросно-ответных диалоговых систем. Существуют различные подходы и принципы построения вопросно-ответных диалоговых систем, но основными являются: на основе метапоисковой системы, на основе системы поиска по аннотированному тексту, на основе экспертной системы, на основе системы поиска в коллекциях вопросов и ответов [13].

Метапоисковая система. В качестве источника данных такая система использует классическую поисковую систему, то есть использует неструктурированные данные, которые делятся на две группы: традиционные неструктурированные документальные и неструктурированные семантические. Система анализирует вопрос пользователя на естественном языке с целью выделить: предположение о семантическом классе ответа, фокус вопроса (вопросительные слова), опора вопроса – остальные члены вопросительного предложения, которые описывают уникальные свойства искомого объекта [14].

Метапоисковая система обычно формулирует запрос по ключевым словам, входящим в опору вопроса. Ключевыми словами можно считать все информативные для частной задачи поиска слова или слова, выделенные по некоторому алгоритму. Результаты поиска обрабатываются компонентами автоматической обработки текста, то есть выделяются все именованные сущности, соответствующие искомому семантическому классу: персоны, географические названия, линейные размеры, названия организаций и др. Далее

синтаксический и семантический разбор позволяют выбрать из всех найденных сущностей наиболее подходящие.

Можно выделить свойства данного подхода: возможность использования доступных инструментов для анализа фрагментов (поиск по ключевым словам, контекстный поиск, полнотекстовый поиск); представление фрагментов в виде графа; вычислительная нагрузка в момент обработки вопроса, связанная с выполнением лингвистических задач; использование неструктурированных данных.

Поиск по аннотированному тексту. Такие системы имеют в своем составе поисковый индекс документов в отличие от метапоисковых. Работают такие системы также с неструктурированными данными. Элементами индекса являются не отдельные слова текста, а объекты детального лингвистического анализа: именованные сущности [15], элементарные синтаксические связки (пары грамматически связанных слов и др.) [4], предикативно-аргументные структуры предложения [16]. Построение индекса происходит с привлечением компьютерной лингвистики: каждый новый документ проходит автоматическую обработку на естественном языке, размечаются объекты вопросно-ответной системы, затем они добавляются в индекс.

Можно выделить свойства данного подхода: меньшая вычислительная нагрузка по сравнению с метапоисковой системой в момент обработки вопроса в реальном времени благодаря специализированному индексу; любые изменения требуют перестроения индекса; использование неструктурированных данных; возможность развернутых ответов.

Экспертная система. В начале 70-х годов прошлого века начинает активно развиваться подход отделения системы работы с правилами – системы вывода и системы хранения самих правил. Теперь информация хранится не в форме данных, а форме знаний – набора простых правил и фактов. А система вывода при помощи объединения знаний из разных правил может получать новую информацию, не хранящуюся в базе знаний системы непосредственно. Подобная концепция получила название *Knowledge Programming,* а системы, которые придерживаются подобного подхода, называют экспертными системами [13].

Основными компонентами экспертной системы являются *база фактов*, *база правил, база автоматически сгенерированных знаний* и *машина вывода*. Современная форма накопления предметного знания представляется как база данных, отображающая ситуационную модель релевантной сферы, то есть профессиональной сферы, для которой предназначена конкретная экспертная система. Экспертная система оперирует не только данными, но и понятийными знаниями, выраженными на естественном языке. Предметное знание – это совокупность сведений о качественных и количественных характеристиках конкретных объектов [13].

База фактов – это структурированная база данных, которая может быть построена автоматически в результате анализа коллекции документов [13].

База правил – формализованные процедуры установления различных типов связей между ними. Основными типами связей являются: *иерархические*, *определяемые отношениями структуризации*, и *семантические связи*, задаваемые *функциональными* и *каузальными* отношениями. Под каузальными связями будем понимать простые отношения причинности, на основе которых можно с некоторой уверенностью считать, что какое-то свойство есть результат действия другого свойства [18].

Функциональные отношения содержат процедурную информацию, позволяющую вычислить одни информационные единицы на основе других, хранящихся в базе фактов. Результатом является база знаний, позиционируемая как семантическая сеть.

Неотъемлемым элементом экспертных систем также является некоторая управляющая структура, которая определяет, какое из правил должно быть проверено следующим. Часто его называют интерпретатором правил или *машиной вывода*.

Можно выделить свойства данного подхода: возможность машинного обучения; способность к адаптации базы правил к домену знаний; необходимость выбирать только

авторитетные исходные тексты для извлечения информации об окружающем мире, однако, эти факты могут противоречить друг другу, и система должна учитывать это; использование структурированных данных.

Поиск в коллекции вопросов и ответов. В социальных системах вопросно-ответного поиска одни пользователи отвечают на вопросы других. Пользователь открывает страницу Web-сайта и формулирует вопрос. Система ищет похожие вопросы в коллекции вопросов и ответов и выдает найденный раздел, где обсуждается вопрос. Если подобный вопрос не существует, создается новый раздел для обсуждения вопроса. На этот вопрос отвечают желающие, а автору 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 [13]. В этой системе необходим модуль извлечения вопросов и ответов из коллекции документов. Вопросно-ответная система непрерывно сканирует все страницы Интернета, анализируя тексты на естественном языке и формулируя возможные вопросы по этому тексту [15]. Кроме того, модуль позволяет поднимать или понижать рейтинг автоматически сгенерированной пары «вопрос-ответ».

Можно выделить свойства данного подхода: возможность развернутых ответов; проверка достоверности ответов другими пользователями; необходимость мотивации пользователей как для пополнения коллекции, так и для оценивания ответов, особенно порожденных автоматически; требуется вспомогательный метод порождения коллекции; использование неструктурированных данных.

Сравнение подходов к построению вопросно-ответных диалоговых систем. В таблице 2 приведены критерии сравнения подходов к построению вопросно-ответных диалоговых систем.

Таблица 2

Критерии сравнения подходов к построению вопросно-ответных диалоговых систем

Критерий	Описание
Структурированные данные	Поддерживает ли подход работу с структурированными данными
Машинное обучение	Возможность машинного обучения
Развернутые ответы	Возможность развернутых ответов на заданный вопрос

В таблице 3 приведено сравнение основных подходов построения вопросно-ответных диалоговых систем по описанным выше критериям.

Таблица 3

Сравнение подходов к построению вопросно-ответных диалоговых систем

Подход	Структурированные данные	Машинное обучение	Развернутые ответы
Метапоисковая система	Нет	Нет	Нет
Поиск по аннотированному тексту	Нет	Нет	Да
Экспертная система	Да	Да	Нет
Поиск в коллекции вопросов и ответов	Нет	Нет	Да

Заключение. Описаны основные подходы к построению вопросно-ответных систем: метапоисковая система; поиск по аннотированному тексту; экспертная система; поиск в коллекции вопросов и ответов. Проведен сравнительный анализ упомянутых подходов по критериям: работа с структурированными данными; возможность машинного обучения; поддержка развернутых ответов. Гибридная вопросно-ответная система может сочетать преимущества разных подходов, в частности, использовать как структурированные, так и неструктурированные данные, выполнять поиск в аннотированном тексте.

Список литературы:

1. Balakrishnan J. Conversational commerce: entering the next stage of AI – powered digital assistants // Annals of Operations Research. – 2021. – № 333. – С. 653 – 687. – DOI: 10.1007/s10479-021-04049-5.

2. Daqar A. The Role of Artificial Intelligence on Enhancing Customer Experience //International Review of Management and Marketing. – 2019. – С. 22 – 31.

3. Дунаев А. Исследовательская система для анализа текстов на естественном языке // Проблемы интеллектуализации и качества систем информатики. Новосибирск.: Институт систем информатики имени А. П. Ершова. – 2006. – С. 55 – 66.

4. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В – М.: МИЭМ, 2011. – 272 с.

5. Митина О. В., Евдокименко А. С. Методы анализа текста: методологические основания и программная реализация // Психология. Психофизиология. – 2010. – 40 (216). – С. 29 – 38.

6. Богуславский И. М., Иомдин Л. Л., Крейдлин Л. Г., Фрид Н. Е., Сагалова И. Л., Сизов В. Г. Модуль универсального сетевого языка (UNL) в составе системы ЭТАП-3 // Труды Международного семинара по компьютерной лингвистике и ее приложениям (Диалог’2000). Протвино, 14–16 июня 2000. / под ред. А. С. Нариньяни. – М.: Изд-во РГГУ, 2000. – Том 2. – С. 48–58.

7. Collins M. Head-Driven Statistical Models for Natural Language Parsing. Computational Linguistics / Computational Linguistics 29 (4), 2003. – С. 589-637.

8. Зайдельман Л.Я., Котов А.А., Зинина А.А., Аринкин Н.А. Система понимания текста для робота Ф-2: синтаксический анализ и извлечение смысла // Восьмая международная конференция по когнитивной науке: Тезисы докладов. Светлогорск, 18–21 октября 2018 г. / Отв. ред. А.К. Крылов, В.Д. Соловьев. – М.: Изд-во «Институт психологии РАН», 2018. – С. 388–391.

9. Аношин П. И. Автоматический анализ текстов. Синтаксический и семантический анализ // Евразийский научный журнал. – 2017. – № 6 – С. 211 – 213.

10. Барышникова Н. Ю. Обработка запросов на естественном языке на основе семантических сетей и шаблонов // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. – 2016. – № 4.

11. Volkova L., Kotov A., Ignatev A. Crowdsourcing-based approbation of communicative behaviour elements on the F-2 robot: perception peculiarities according to respondents // Samsonovich A.V., Liu T. (eds.) Biologically Inspired Cognitive Architectures 2023. Studies in Computational Intelligence, vol. 1130. – Cham, Switzerland: Springer Nature, 2024. – С. 932–945.

12. Кан Д. А. Задача синтеза предложений на естественном языке // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. – 2009. – № 3. – С. 205 – 212.

13. Черноморова Т. С., Воробьев С. П. Классификация и принципы построения систем вопросно-ответного поиска // Бюллетень науки и практики. – 2020. – Т. 6, № 8. – С. 145 – 156.

14. Соловьёв А., Пескова О. Построение вопросно-ответной системы для русского языка: модуль анализа вопросов // Новые информационные технологии в автоматизированных системах. – 2010. – № 13. – С. 41 – 49.

15. Experiments with interactive question-answering / S. Harabagiu [и др.] // Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL’05). – 2005. – С. 205 – 214.

16. Gonz´alez J. L. V., Rodr´ıguez A. F. A Semantic Approach to Question Answering Systems // Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000. Т. 500 – 249 / под ред. E. M. Voorhees, D. K. Harman. – National Institute of Standards, Technology (NIST), 2000. – (NIST Special Publication).

17. Minsky M. A framework for representing knowledge // Winston, P.H. (ed.) The Psychology of Computer Vision. – New York: McGraw-Hill, 1975. – С. 211–277.

18. Котов А. А. Особенности каузального мышления у экспертов и новичков // Когнитивная психология: феномены и проблемы. – М.: ЛЕНАНД, 2014. – С. 87 – 107.

19. Ferrucci D. Building Watson: An overview of the DeepQA project // AI magazine. – 2010. – Т. 31, № 3. – С. 59 – 79.