Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ «Информатика и системы управления»

КАФЕДРА ИУ-7 «Программное обеспечение эвм и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Классификация известных методов обработки текстов в вопросно-ответных системах»

Студент **ИУ7-53Б** _____ **Завойских Е. В.**

Руководитель _____ **Волкова Л. Л.**

2022 г.

РЕФЕРАТ

Расчетно-пояснительная записка 28 с., 1 рис., 6 табл., 16 ист., 1 прил.

Ключевые слова: анализ вопроса, информационный поиск, извлечение потенциальных ответов, валидация ответов, метод N-грамм, использование шаблонов.

Научно-исследовательская работа представляет собой обзор существующих методов обработки текстов в вопросно-ответных системах, формулировку критериев сравнения методов и классификацию этих методов по критериям.

Содержание

Введение

В современном мире стремительно растет объем доступной информации. В частности, огромное количество данных, в том числе неструктурированных, доступно в сети Интернет. Так как значительная часть информации представлена в виде текстов на естественном языке, возникла необходимость анализа и обработки таких текстов. Одной из основных задач этого направления является обработка текстов в вопросно-ответных системах. В последнее время интерес исследователей смещается от традиционного поиска по запросу в сторону интеллектуальных систем поиска информации, поэтому существует множество методов, решающих задачу обработки текстов в вопросно-ответных системах.

Целью данной работы является классификация известных методов обработки текстов в вопросно-ответных системах.

Для достижения поставленной цели необходимо решить следующие задачи:

— провести обзор предметной области вопросно-ответного поиска;

— рассмотреть основные этапы работы вопросно-ответных систем;

— описать существующие методы обработки текстов, относящиеся к каждому из этапов;

— сформулировать критерии сравнения описанных методов;

— сравнить методы по сформулированным критериям.

1 Анализ предметной области

1.1 Вопросно-ответный поиск

Современные поисковые системы выдают информацию в виде документов, веб-страниц, изображений и видео по запросу, но распознавать запросы на естественном языке и формировать ответы в соответствующем формате большая часть из них не способна. В результате подобных запросов пользователь получается множество ссылок на различные сайты и документы, которые ему необходимо изучить, чтобы найти ответ на свой вопрос. От обычных поисковых систем выгодно отличаются системы вопросно-ответного поиска.

Вопросно-ответные системы — это вид информационно-поисковых систем, способных обрабатывать введенный пользователем вопрос на естественном языке и выдавать осмысленный ответ [1].

В системах вопросно-ответного поиска активно используются технологии обработки естественного языка. Сначала на вход системе подаётся запрос, сформулированный в виде вопросительного предложения на естественном языке. Далее этот запрос обрабатывается, происходит поиск и вывод ответа в виде одного или нескольких слов на естественном языке, либо небольшого фрагмента текста, сформированного системой в результате анализа и обработки разнообразных источников данных. Источником информации для поиска ответа может быть как Интернет, так и локальное хранилище данных. Вопросно-ответные системы отличаются тематикой вопросов, которые они могут обрабатывать.

1.2 Виды вопросно-ответных систем

Системы вопросно-ответного поиска можно разделить на две группы [2]:

1) системы без специализации тематики;

2) узкоспециализированные по тематике системы.

Первый вид вопросно-ответных систем ориентирован на обработку вопросов по любым предметным областям. В качестве источника информации общие системы используют большой корпус документов или сеть Интернет.

Узкоспециализированные системы направлены на ответы на вопросы по конкретным предметным областям, например, медицина, юриспруденция. Релевантность и полнота ответов зависят от полноты знаний о домене, которому посвящён диалог.

Существуют различные принципы построения вопросно-ответных систем, но основными являются следующие [1]:

1) метапоисковые вопросно-ответные системы;
2) вопросно-ответные системы поиска ответа по аннотированному тексту;
3) вопросно-ответные системы поиска ответа в коллекциях вопросов и ответов;
4) экспертные вопросно-ответные системы.

1.2.1 Метапоисковые системы

В качестве источника данных такая система использует классическую поисковую систему, то есть использует неструктурированные данные.

Вопросно-ответная система после получения на вход от пользователя вопросительного предложения на естественном языке обрабатывает это предложение и формирует запрос для поисковой системы из ключевых слов, которые выбираются исходя из самого вопросительного предложения. Результаты поиска обрабатываются существующими компонентами систем автоматической обработки текста. Например, выделяются все именованные сущности, соответствующие искомому классу ответа: персоны, географические названия, названия организаций, линейные размеры и другие. Далее синтаксический и семантический разбор позволяют выбрать из всех найденных сущностей наиболее подходящие.

1.2.2 Системы поиска ответа по аннотированному тексту

Такие системы имеют в своем составе поисковый индекс документов и работают с неструктурированными данными. Элементами индекса являются не отдельные слова текста, а объекты детального лингвистического анализа: именованные сущности, элементарные синтаксические связки (пары грамматически связанных слов и другие). Построение индекса происходит с привлечением компьютерной лингвистики, а именно каждый новый документ проходит этапы автоматической обработки текста на естественном языке, размечаются объекты вопросно-ответной системы (именованные сущности, элементарные синтаксические связки), затем они добавляются в индекс.

1.2.3 Системы поиска ответа в коллекциях вопросов и ответов

В социальных системах вопросно-ответного поиска (англ. *collaborative question answering*) одни пользователи отвечают на вопросы других. Пользователь открывает страницу веб-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.

1.2.4 Экспертные системы

Вопросно-ответные системы, построенные по принципу работы со структурированными базами данных, можно отнести к классу экспертных систем.

Основными компонентами экспертной системы являются база фактов и база правил. База фактов — это структурированная база данных, которая может быть построена автоматически в результате анализа коллекции докумен-

тов. Этот процесс аналогичен построению аннотированного индекса. Однако он происходит на более детальном уровне обработки естественного текста: извлекаются не синтаксические конструкции, а факты. В базе правил хранятся процедуры для установления различных типов связей между фактами. Эти процедуры содержат информацию, позволяющую выполнять логический вывод новой информации на основе имеющихся фактов. Результатом является база знаний, позиционируемая как семантическая сеть.

Важным элементом экспертных систем также является некоторая управляющая структура, которая определяет — какое из правил должно быть проверено следующим при формировании новой информации.

Такие системы являются узкоспециалированными из-за сложности организации базы фактов. Подобная база должна состоять только из достоверной информации о предметной области.

1.3 Этапы вопросно-ответного поиска

Процесс работы вопросно-ответной системы можно разделить на следующие этапы [3]:

1) этап анализа вопроса;
2) этап информационного поиска;
3) этап извлечения потенциальных ответов;
4) этап валидации ответов.

1.3.1 Анализ вопроса

На этапе анализа вопроса происходит ввод пользователем вопроса на естественном языке и дальнейшая его обработка. Вопросы можно разделить по виду ответа на следующие [4]:

— фактографические вопросы;
— вопросы причины;

— вопросы мнения.

Фактографический вопрос — это вопрос о различных сведениях без их анализа и обобщения, ответ на данный вопрос обычно краток. Примерами фактографических вопросов являются вопросы о персонах, о времени, вопросы, требующие ответа «да» или «нет».

Вопросы причины требуют логического анализа текста, определения между предложениями причинно-следственной связи и являются самыми сложными в плане нахождения ответа. Существующие решения в области анализа текста и вывода его логической структуры довольно плохо справляются с данным видом вопросов.

Вопросы мнения предусматривают собой поиск и глубокий анализ блогов и различных сайтов СМИ.

На этапе анализа вопроса ставится следующая задача: для вопроса на естественном языке выделить фокус вопроса, опору вопроса и определить семантический тэг ответа [5].

Фокус вопроса — это такие сведения, содержащиеся в вопросе, которые несут в себе информацию об ожиданиях пользователя от информации в ответе, т. е. вопросительные слова, обозначающие искомую информацию, например, «в каком городе», «кто», «в каком году», «сколько», «какого цвета» и другие.

Опора вопроса — это остальная часть вопроса (после «вычета» фокуса), которая несёт в себе информацию, поддерживающую выбор конкретного ответа. В опору вопроса входят ключевые слова, по которым система формулирует запрос для информационного поиска.

Семантический тэг ответа — класс вопроса, соответствующий типу возможного ответа. Многие вопросно-ответные системы имеют встроенную систему поддерживаемых типов ожидаемого ответа, иногда используется иерархическая структура представления в виде таксономии в зависимости от дальнейшей стратегии поиска и извлечения ответа.

1.3.2 Информационный поиск

На этом этапе производится поиск релевантных запросу документов, а также получение текстовых фрагментов, содержащих ответ [3]. Результат дол-

жен содержать необработанный или в редких случаях готовый ответ на вопросительное предложение пользователя.

1.3.3 Извлечение потенциальных ответов

На данном этапе распознаются и извлекаются из полученных текстовых фрагментов потенциальные ответы на вопрос. Важную роль в выделении ответа играет тип ответа.

1.3.4 Валидация ответов

На последнем этапе производится анализ списка кандидатов, все кандидаты оцениваются и выбирается наиболее подходящий.

2 Описание существующих решений

В данном разделе представлено краткое описание существующих методов обработки текстов на каждом этапе работы вопросно-ответной системы.

2.1 Этап анализа вопроса

2.1.1 Символьные шаблоны вопросов

Одним из способом определить тэг или фокус в вопросе является подготовка шаблонов (регулярных выражений) для распознавания распространенного вопросительного оборота [5]. В таблице 2.1 приведены некоторые правила, используемые в системе OpenEphyra для английского языка [6]:

Таблица 2.1 – Символьные шаблоны вопросов из системы OpenEphyra

Семантический тэг	**Регулярное выражение вопроса**
Date → Weekday	(what|which|name) (.*)?(day of (the)?week|weekday)
Location → Country	(what|which|name) (.*)?(colony|country|nation)
Size → Length	how (deep|far|high|long|tall|wide)
Size → Length	(how large in|how many) (foot|inch|.*meter|mile|yard)

Недостатки такого подхода следующие.

1) Невозможность покрыть большую часть реальных вопросов пользователей. Набор вопросов подбирается так, чтобы обработать конкретный набор тестовых заданий. Выйти за пределы этого покрытия «неудобным вопросом» достаточно легко.

2) Связь между вопросительными словами и семантическими тэгами не так прямолинейна. Так, слово «кто» может подразумевать персону, организацию, страну или народ (например, в вопросе «Кто захватил Константинополь?»).

2.1.2 Синтаксические шаблоны вопросов

Для выделения фокуса вопроса следующим шагом после символьных шаблонов стал метод синтаксических шаблонов. Данные шаблоны значительно сложнее символьных, однако могут использоваться в широком круге вопросов. В основе метода лежит предположение, что фокус вопроса часто находится в определённом синтаксическом отношении с вопросительным словом, может быть не в одном, но набор вариантов этих отношений ограничен [5].

Сначала выполняется синтаксический разбор предложения, в результате чего получается синтаксическое дерево. Далее это дерево вопроса сравнивается с синтаксическим шаблоном для распознавания фокуса и, в случае совпадения, фокусом считаются члены предложения, соответствующие позиции фокуса в шаблоне.

2.1.3 Статистика употребления слов в вопросах

Это метод автоматического обучения статистической модели для простановки семантического тэга [5]. Для каждого вопроса из обучающей выборки выделяют три «потока» признаков:

— все слова как есть и дополнительные метки к некоторым из них;

— метки частей речи слов и порядковые номера слов в предложении;

— фокусные слова с гиперонимами, согласно лексическому тезаурусу.

После разметки вручную коллекции из вопросов подсчитывается, какие свойства чаще означают каждый семантический тэг. Для этого используется математический аппарат максимизации энтропии. Метод максимума энтропии [7] является вероятностным классификатором, который основан на принципе максимальной энтропии. По данному принципу распределения вероятности являются равномерными (имеют максимальную энтропию), если нет оснований считать иначе.

Недостатком статистического метода является необходимость создания большой обучающей коллекции вопросов вручную.

2.2 Этап информационного поиска

2.2.1 Деление на абзацы

Наиболее простым способом выделения текстовых фрагментов из набора документов является деление текста на абзацы. Затем выбираются те фрагменты, которые содержат наибольшее количество ключевых слов.

2.2.2 Использование окон параграфа

В вопросно-ответной системе, разработанной в Южном Методистском Университете США [8], был использован другой способ получения фрагментов документов для последующего извлечения из них ответа.

Например, у нас есть набор ключевых слов $k1$, $k2$, $k3$, $k4$ и в параграфе текста $k1$ и $k2$ встречаются каждое дважды, тогда как $k3$ — только один раз, и $k4$ ни разу. Далее вводится понятие окна параграфа — это фрагмент, содержащий весь текст между ключевым словом с самой низкой позицией в окне и ключевым словом с самой высокой позицией в окне. В данном примере будет четыре разных окна, определяемых ключевыми словами: [$k1$–соответствие$_1$, $k2$–соответствие$_1$, $k3$], [$k1$–соответствие$_2$, $k2$–соответствие$_1$, $k3$], [$k1$–соответствие$_1$, $k2$–соответствие$_2$, $k3$] и [$k1$–соответствие$_2$, $k2$–соответствие$_2$, $k3$].

Для каждого окна параграфа вычисляются следующие оценки:

— $Same\_word\_sequence\_score$ содержит количество слов из вопроса, которые распознаются в той же последовательности в текущем окне;

— $Distance\_score$ содержит количество слов, разделяющих самые удаленные ключевые слова в окне;

— $Missing\_keywords\_score$ содержит количество не попавших в окно ключевых слов.

Далее выполняется сортировка окон параграфов по трем различным показателям:

— наибольшее значение $Same\_word\_sequence\_score$;

— наибольшее значение $Distance\_score$;

— наименьшее значение $Missing\_keyword\_score$.

2.3 Этап извлечения потенциальных ответов

2.3.1 Использование шаблонов

Для каждого типа ответа составляются шаблоны, с помощью которых в текстовых фрагментах производится поиск и выделение кандидата ответа [9]. Для выбора ответа используется информация о типе ожидаемого ответа, полученная на этапе анализа вопроса, и символьные шаблоны. Шаблоны можно создавать как вручную, так и автоматическими обучаемыми алгоритмами.

2.3.2 Использование N-грамм

N-грамма – это последовательность из N слов, идущих в каком-то тексте подряд. На первом этапе из фрагмента текста извлекаются униграммы, биграммы и триграммы. Каждой из них назначается вес, соответствующий количеству текстовых фрагментов, в которых она нашлась. Далее для каждой N-граммы вес корректируется с учетом типа ожидаемого ответа. После N-граммы ранжируются, выбираются N-граммы с наиболее высокими оценками и конструируется ответ путем объединения перекрывающихся N-грамм [10].

2.4 Этап валидации ответов

2.4.1 Пересечение множеств слов и/или грамматических отношений

В данном методе применяется модель мешка слов (англ. *bag of words*) [11]. Мера «подтверждения» ответа на вопрос вычисляется по формуле (2.1):

$$E = \frac{|Q \cap T|}{|Q|}, \tag{2.1}$$

где Q — множество слов в вопросе, T — множество слов во фрагменте текста, содержащем потенциальный ответ.

Усложнением является использование не множества слов, а множества синтаксических отношений, т.е. пар слов, связанных грамматической связью: $R(N1, N2)$ [12].

2.4.2 Сопоставление сказуемых

Ответ и фрагмент текста с потенциальным ответом предварительно проходят аннотацию семантическими ролями. В результате слова предложения получают метку либо сказуемого, либо аргумента при каком-то сказуемом. Сравниваются два сказуемых со всеми зависимыми словами: одно сказуемое из вопроса, другое — из текстового фрагмента. Схожесть двух сказуемых вычисляется как произведение лексической схожести глаголов (например, расстояние по словарю WordNet) и схожести наборов аргументов:

$$Sim_{Pred} = Sim_{Verb} \cdot Sim_{Args}. \tag{2.2}$$

Схожесть аргументов предиката вопроса p_q и предиката p_a из фрагмента

текста вычисляется следующим образом:

$$Sim_{Args}(p_a, p_q) = \frac{\sum\limits_{t_a \in T_a} \max\limits_{t_q \in T_q}(Sim_{ExpTerm}(t_a, t_q))}{|T_q| + |\{t_a \in T_a | \max\limits_{t_q \in T_q}(Sim_{ExpTerm}(t_a, t_q)) = 0\}|}. \quad (2.3)$$

Мера схожести двух термов вычисляется по формуле (2.4):

$$Sim_{Term}(t_1, t_2) = \frac{|W1 \cap W2|}{|W1 \cup W2|}, \quad (2.4)$$

где $W1$ и $W2$ — множества контекстных слов из описания значения терма в словаре.

Если в вопросе или фрагменте текста несколько сказуемых, то вычисляется схожесть всех пар. Наибольшее из полученных чисел и будет мерой подтверждения ответа фрагментом текста [12].

2.4.3 Расстояние редактирования для деревьев

Далее рассматривается задача вычисления схожести деревьев грамматических зависимостей между словами двух предложений: вопросительного и повествовательного [12]. Для предложения-кандидата и вопросительного предложения строятся так называемые деревья грамматических зависимостей, отражающие связи между словами, рассматривая их как части речи. Пример дерева для вопросительного предложения приведен на рисунке 2.1 [13].

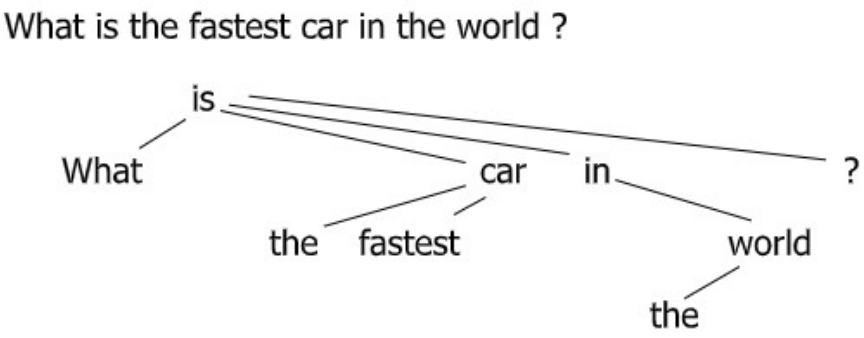

Рисунок 2.1 – Пример дерева грамматических зависимостей для вопроса

В данном методе применяется естественная метрика схожести деревьев: минимальное число операций редактирования, необходимых для трансформа-

ции одного графа в другой. Доступные операции редактирования: удаление вершины, вставка, замена. Каждая такая операция s_i имеет некоторый вес $y(s_i)$.

Пусть $S =< s1; s2; \ldots; sk >$ — последовательность операций, приводящая к трансформации дерева предложения-кандидата на ответ в дерево вопроса. Тогда стоимость этой трансформации есть сумма стоимостей операций.

$$y(S) = \sum y(s_i). \tag{2.5}$$

Последовательность с наименьшей стоимостью операций редактирования вычисляется по формуле (2.6):

$$\delta(T_p, T_q) = \min_S(y(S)|S(T_p) = T_q). \tag{2.6}$$

В итоге выбираются кандидаты с наименьшей оценкой.

3 Классификация существующих решений

В данном разделе предлагаются критерии оценки методов и проводится сравнение по выделенным критериям.

3.1 Сравнение и оценка для этапа анализа вопроса

Для оценки методов данного этапа выбраны такие критерии, как сложность реализации, влияющий на скорость разработки модуля анализа вопросов, покрытие вопросов методом и необходимость предварительной обработки вопроса.

Результаты сравнения методов анализа вопроса приведены в таблице 3.1. Оценки сложности реализации и покрытия даны по [5].

Таблица 3.1 – Сравнение методов анализа вопроса

Метод	Сложность реализации	Покрытие вопросов	Необходимость преварительной обработки вопроса
Символьные шаблоны	Низкая	Низкое	Нет
Синтаксические шаблоны	Средняя	Высокое	Есть
Статистика употребления слов	Высокая	Среднее	Нет

Метод символьных шаблонов обладает излишней простотой, поэтому корректно работает лишь в ограниченном числе случаев. Синтаксические шаблоны способны покрыть значительную часть реальных вопросов, но их сложнее подготовить, так же требуется построить синтаксическое дерево вопроса. Для корректной работы статистического метода необходимо создание большой обучающей коллекции вопросов вручную.

Представляется целесообразным последовательное использование всех

методов. Символьные шаблоны будут эффективны на первом этапе обработки, и в случае полного соответствия вопроса шаблону обработка прекращается. В противном случае подключается имеющаяся статистика, и вопрос анализируется согласно ей. Если же для текущего вопроса статистика не собрана или недостаточно достоверна, применяются наиболее общие синтаксические шаблоны.

3.2 Сравнение и оценка для этапа информационного поиска

Для оценки методов данного этапа выбраны такие критерии, как учет методом количества ключевых слов в текстовом фрагменте, учет порядка следования ключевых слов и учет расстояния между ключевыми словами [4].

Результаты сравнения методов приведены в таблице 3.2.

Таблица 3.2 – Сравнение методов на этапе информационного поиска

Метод	Учет числа ключевых слов	Учет порядка ключевых слов	Учет расстояния между словами
Деление на абзацы	Есть	Нет	Нет
Использование окон параграфа	Есть	Есть	Есть

По данным таблицы 3.2 наиболее точным является метод, использующий окна параграфа, так как он учитывает не только количество ключевых слов в текстовом фрагменте, но и их расположение.

3.3 Сравнение и оценка для этапа извлечения потенциальных ответов

Для оценки методов используются следующие метрики:

— среднеобратный ранг (MRR) — оценка извлеченных ответов, упорядоченных по вероятности и правильности;

— доля вопросов, на которые были даны правильные ответы.

Результаты сравнения методов извлечения потенциальных ответов приведены в таблице 3.3.

Таблица 3.3 – Сравнение методов извлечения потенциальных ответов

Метод	MRR	Доля вопросов
Использование шаблонов [14]	0.29	0.25
Использование N-грамм [15]	0.42	0.49

По данным таблицы 3.3 наиболее точным является метод извлечения ответа с использованием N-грамм, однако важно отметить, что значения в первой строке таблицы 3.3 сильно зависят от полноты шаблонов и могут отличаться в различных работах.

3.4 Сравнение и оценка для этапа валидации ответов

Данный способ оценки валидации ответов основан на традиционном подходе к оценке в задаче классификации [16]. Задача валидации рассматривается как задача бинарной классификации: тройку <вопрос, ответ, текстовый фрагмент> требуется отнести к одному из классов — верный ответ или неверный.

В таблице 3.4 приведены четыре возможных исхода решения задачи классификации.

Таблица 3.4 – Категории результата классификации ответов

Наблюдаемый результат	Ожидаемый результат	
	Верный ответ	Неверный ответ
Верный ответ	tp (true-positive)	fp (false-positive)
Неверный ответ	fn (false-negative)	tn (true-negative)

На основе этой таблицы определяются традиционные метрики качества классификации:

Accuracy — доля ответов, по которым классификатор принял правильное решение. Вычисляется по формуле (3.1):

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}. \tag{3.1}$$

Точность (precision) — доля ответов, действительно принадлежащих данному классу, относительно всех ответов, которые система отнесла к этому классу. Вычисляется по формуле (3.2):

$$Precision = \frac{tp}{tp + fp}. \tag{3.2}$$

Полнота (recall) — доля ответов, причисленных классификатором к данному классу, относительно всех ответов, принадлежащих ему в тестовой выборке. Вычисляется по формуле (3.3):

$$Recall = \frac{tp}{tp + fn}. \tag{3.3}$$

F-мера — среднее гармоническое точности и полноты, вычисляется по формуле (3.4):

$$F_\beta = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{3.4}$$

где коэффициент β может рассматриваться как относительная степень важности показателей полноты и точности. При значении коэффициента равном 1/2 точность вдвое важнее полноты, при значении равном 2 полнота вдвое важнее точности.

Результаты сравнения методов валидации ответов приведены в таблице 3.5. Значения метрик даны по [16].

Таблица 3.5 – Сравнение методов валидации ответов

Метод	$Accuracy$	$Precision$	$Recall$	$F_{0.5}$
Пересечение множеств слов	0.62	0.51	0.47	0.48
Сопоставление сказуемых	0.70	0.72	0.27	0.54
Расстояние редактирования	0.64	0.45	0.09	0.26

По данным таблицы 3.5, наиболее точным является метод сопоставления сказуемых. Хотя метод пересечения множеств слов имеет лучший показатель полноты, на данном этапе показатель точности является более важным.

Заключение

В рамках научно-исследовательской работы была проведена классификация известных методов обработки текстов в вопросно-ответных системах.

В результате сравнения методов были выделены метод, использующий окна параграфа, для этапа информационного поиска, метод извлечения ответа с использованием N-грамм и метод сопоставления сказуемых для этапа валидации ответа, как наиболее точные из представленных. Для этапа анализа вопроса было определено целесообразным последовательное использование трех методов анализа.

В итоге, в ходе данной работы решены все задачи:

— проведен обзор предметной области вопросно-ответного поиска;

— рассмотрены основные этапы работы вопросно-ответных систем;

— описаны существующие методы обработки текстов, относящиеся к каждому из этапов;

— сформулированы критерии оценки сравнения описанных методов;

— проведено сравнение методов по сформулированным критериям.

Таким образом, поставленная цель достигнута.

Список использованных источников

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

2. Рыбак, К. В. Обзор современного состояния интеллектуальных вопросно-ответных систем / К. В. Рыбак, А. В. Кошкаров // Вестник науки. — 2020. — Т. 1, № 6. — С. 202–205.

3. Деревянко, Д. В. Формальные методы разработки вопросно-ответной системы на естественном языке / Д. В. Деревянко, Д. Е. Пальчунов // Вестник Новосибирского государственного университета. Серия: Информационные технологии. — Новосибирск: НГУ, 2014. — Т. 12, № 3. — С. 34–47.

4. Allam, A. M. N. The Question Answering Systems: A Survey / A. M. N. Allam, M. H. Haggag // International Journal of Research and Reviews in Information Science (IJRRIS). — 2012. — Vol. 2, № 3. — P. 10–21.

5. Соловьев, А. А. Построение вопросно-ответной системы для русского языка: модуль анализа вопросов / А. А. Соловьев, О. В. Пескова // Новые информационные технологии в автоматизированных системах: материалы 13-го научно-практического семинара. — М.: МИЭМ, 2010. — С. 41–49.

6. Schlaefer, N. The Ephyra QA system at TREC 2006 / N. Schlaefer, P. Gieselmann, G. Sautter // Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006. — Gaithersburg, 2006. — P. 34–47.

7. Gritta M. A Comparison of Techniques for Sentiment Classification of Film Reviews // CoRR. — 2019. — URL: `http://arxiv.org/abs/1905.04727`.

8. Moldovan, D. The Structure and Performance of an Open-Domain Question Answering System / D. Moldovan, S. Harabagiu, M. Pasca [et al.] // ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. — Hong Kong, 2000. — P. 563–570.

9. Ravichandran, D. Learning Surface Text Patterns for a Question Answering System / D. Ravichandran, E. Hovy // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic. — 2002. — P. 41–47.

10. Dumais, S. Web question answering: is more always better? / S. Dumais, M. Banko, E. Brill [et al.] // SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. — New York, 2002. — P. 291–298.

11. Попов, В. В. Естественный текст: математические методы атрибуции / В. В. Попов, Т. В. Штельмах // Вестник Волгу. Серия 2: Языкознание. — Волгоград: ВолГУ, 2019. — Т. 18. — С. 147–158.

12. Соловьёв, А. А. Алгоритмы валидации ответов в задаче вопросно-ответного поиска / А. А. Соловьёв // Вестник ВГУ. Серия: Системный анализ и информационные технологии. — Воронеж: ВГУ, 2011. — № 2. — С. 181–188.

13. Punyakanok, V. Natural language interface via dependency tree mapping: An application to question answering / V. Punyakanok, D. Roth, W. Yih // AI and Math. — 2004. — P. 22–34.

14. Echihabi, A. How To Select An Answer String? / A. Echihabi, U. Hermjakob, E. Hovy [et al.] // Advances in Open Domain Question Answering. — Dallas, 2006. — P. 383–406.

15. Radev, D. R. Probabilistic question answering on the web / D. R. Radev, W. Fan, H. Qi [et al.] // The 11th International World Wide Web Conference. — Honolulu: ACM, 2002. — P. 408–419.

16. Соловьёв, А. А. Взвешенная погрешность — новая метрика для оценки качества валидации ответов в задаче вопросно-ответного поиска / А. А. Соловьёв // Вестник МГТУ им. Н.Э. Баумана. Серия «Приборостроение». — М.: МГТУ им. Н. Э. Баумана, 2013. — № 3. — С. 58–64.

Приложение А

В графическую часть научно-исследовательской работы входит презентация.