**Система ответов на вопросы AnswerBus**

Чжипин Чжэн

Школа информации

Мичиганский университет анн-Арбор, MI 48109 zzheng@umich.edu

# АБСТРАКТНЫЙ

AnswerBus - это система ответов на вопросы с открытым доменом, основанная на веб-поиске информации на уровне предложений. Он принимает вопросы пользователей на естественном языке на английском, немецком, французском, испанском, итальянском и португальском языках и предоставляет ответы на английском языке. Пять поисковых систем и каталогов используются для извлечения веб-страниц, которые имеют отношение к вопросам пользователей. Из веб-страниц AnswerBus извлекает предложения, которые содержат ответы. В настоящее время показатель правильных ответов на 200 вопросов TREC-8 составляет 70,5%, а среднее время ответа на вопросы составляет семь секунд. Производительность AnswerBus с точки зрения точности и времени отклика лучше, чем у других подобных систем. [[1]](#footnote-1)

# Ключевые слова

ответы на вопросы в открытом домене, словарь QA, поиск информации

# ЗНАКОМСТВО

Исследования автоматизированных ответов на вопросы (QA) могут быть датированы 1960-ми годами, тем не менее, часто ограничивались предметно-ориентированными экспертными системами ([12], [9]). Исследователи экспериментировали с системами контроля качества, основанными на закрытых, предварительно помеченных корпусах (например, [8], [13], [3]) или базах знаний (например, [11], [5], [4]). Многие из этих систем сосредоточены на задачах Text REtrieval Conference (TREC). Исследователи также пытались построить системы контроля качества на больших коллекциях документов.

в Интернете путем сочетания извлечения информации и самых передовых технологий поиска информации (например, [7], [12], [1]). В последнее время исследователей привлекла задача разработки систем контроля качества с открытым доменом на основе коллекций документов реального мира, особенно Всемирной паутины. Примеры таких систем включают LCC([7]), QuASM[[2]](#footnote-2)[[3]](#footnote-3), IONAUT([1]), START([11]) и Webclopedia([10]). На данном этапе эти системы, как правило, имеют длительное время отклика и/или показатели точности, которые могут быть неприемлемыми для пользователей. [[4]](#footnote-4)[[5]](#footnote-5)[[6]](#footnote-6)

AnswerBus - это система ответов на вопросы с открытым доменом, основанная на поиске информации на уровне предложения. Он принимает вопросы пользователей на естественном языке на английском, немецком, французском, испанском, итальянском и

Португальский язык и извлекает возможные ответы из Интернета. Он может отвечать на вопросы пользователей в течение нескольких секунд. Пять поисковых систем и каталогов (Google, Yahoo, WiseNut, AltaVista и Yahoo News) используются для извлечения веб-страниц, которые потенциально содержат ответы. Из веб-страниц AnswerBus извлекает предложения, которые содержат ответы. В настоящее время показатель правильных ответов на 200 вопросов TREC-8 составляет 70,5%. AnswerBus демонстрирует, что практические ответы на вопросы в Интернете вполне осуществимы.

На рисунке 1 описан рабочий процесс AnswerBus.

AnswerBus принимает вопрос пользователя на естественном языке.

Простой модуль распознавания языка определит, на английском ли вопросе или на любом из других пяти языков. Если язык вопроса не английский, AnswerBus отправит оригинальный вопрос и языковую информацию о вопросе в инструмент перевода AltaVista BabelFish и получит вопрос, который был переведен на английский язык. [[7]](#footnote-7)

Переведенный вопрос

Тип вопроса

Хит-листы из поисковых систем

Поисковая система

Конкретный запрос

Выбранные поисковые системы

Извлеченные предложения

Ранжированные ответы

Ответ Кандидатам

Вопрос пользователя

Совпадающие слова

Рисунок 1 Рабочий процесс AnswerBus

Остальная часть процесса состоит в основном из четырех этапов: 1) выбрать две или три поисковые системы из пяти для поиска информации и сформировать запросы конкретной поисковой системы на основе вопроса; 2) связаться с поисковыми системами и получить документы, указанные в верхней части хит-листов; 3) извлекать из документов предложения, потенциально содержащие ответы; 4) ранжируйте ответы и возвращайте лучшие варианты с контекстными URL-ссылками на пользователя. Вместо того, чтобы возвращать фрагмент текста фиксированной длины, AnswerBus возвращает предложения, которые могут предоставить пользователям некоторую контекстную информацию для ответов.

# ПОИСК СООТВЕТСТВУЮЩИХ ДОКУМЕНТОВ

AnswerBus стремится получить достаточно релевантных документов из поисковых систем в течение времени отклика, приемлемого для пользователей. Основными задачами на этом этапе являются выбор одной или нескольких подходящих поисковых систем для конкретного вопроса пользователя, а затем формирование запросов, которые адаптированы к вопросу, а также выбранным поисковым системам. Формирование запросов является важной процедурой, поскольку она может в значительной степени влиять на запоминание и точность ответов на вопросы и скорость работы системы.

## Выбор поисковой системы

Различные поисковые системы или каталоги могут подходить для разных типов вопросов. Например, для вопросов о текущих событиях Yahoo News может быть лучшим выбором, чем Google. Таким образом, для конкретного вопроса AnswerBus выбирает использование двух-трех наиболее подходящих поисковых систем среди пяти.

Чтобы определить, какие поисковые системы лучше всего подходят для конкретного вопроса, AnswerBus предварительно ответил на 2000 образцов вопросов, включая вопросы TREC 8 и TREC 9, а также вопросы, введенные тестирующими пользователями. Он отправил запросы, основанные на этих вопросах, во все пять поисковых систем. Для каждого вопроса он записывал количество возможных ответов, которые пришли из разных поисковых систем. Все слова, используемые в запросах, индексируются. Например, для Word 1 Google может вернуть 8 ответов, AltaVista возвращает 4 ответа, а Yahoo возвращает 7 ответов; для Word 2, Google 6 ответы, AltaVista 6 ответы и

Yahoo 5 ответы. Для запроса с помощью Word 1 и Word

2 AnswerBus выберет Google (8+6) и Yahoo (7+5). Если запрос содержит слова, не включенные в индексированный список, AnswerBus использует среднюю доходность поисковых систем для всех проиндексированных слов, чтобы определить, какие поисковые системы наиболее подходят. **2.2 Формирование запросов для поисковых систем**

Большинство поисковых систем не предназначены для вопросов на естественном языке. [2] показывает, что системы контроля качества, использующие хорошие запросы, значительно превзойдут базовые веб-поисковые системы и коммерческую поисковую систему, специализирующуюся на ответах на вопросы. [6] утверждает, что использование дополнительной информации для управления процессом поиска дает более ценные результаты, чем рассмотрение только запроса.

Запрос, специфичный для поисковой системы, здесь относится к запросу, сформированному из вопроса пользователя на естественном языке, который относится к конкретным поисковым системам, и будет генерировать оптимальные результаты поиска. «Оптимальный» не обязательно означает отзыв наибольшего числа соответствующих документов. Вместо этого это означает наилучшие результаты с точки зрения как отзыва документов, так и времени на извлечение документов для системы контроля качества. Например, пользователь хочет узнать «Насколько высока гора Эверест». Если полный вопрос отправляется в коммерческую поисковую систему, такую как Google, обычно пользователь сможет собрать документы, содержащие ответ. Тем не менее, больше веб-страниц в возвращенном списке попаданий не будет иметь отношения к вопросу, что приведет к снижению точности. Конечно, требуется гораздо больше времени для извлечения и обработки документов. Таким образом, необходимо сформировать *специфический поисковый запрос*, в данном случае «Гора Эверест», который наилучшим образом подойдет поисковой системе и даст оптимальные результаты поиска.

Тем не менее, задача формирования конкретного запроса поисковой системы не всегда так проста, как в этом примере, учитывая, что каждый вопрос имеет свою уникальную структуру и содержание, в то время как разные поисковые системы имеют разные правила в отношении запросов, которые они принимают.

Некоторые подходы расширяют один запрос до нескольких запросов, например, добавляя синонимы, затем либо объединяют эти запросы вместе с оператором "OR", либо сохраняют их как отдельные запросы. Эти подходы могут увеличить возможность получения правильных ответов, однако могут также усложнить поиск и значительно продлить время отклика на поиск, и в целом могут нанести ущерб результатам QA.

Скорость вычислений считалась очень важной на протяжении всей разработки AnswerBus. Для масштабируемой веб-системы контроля качества необходимо найти компромисс между скоростью и отзывом ответов. Таким образом, AnswerBus не пытается найти лучший запрос; вместо этого он пытается найти достаточно хороший запрос, который будет выполнять задачу поиска очень быстро. Основное внимание было уделено генерации одного простого запроса вместо расширенного.

Для формирования запросов сочетается несколько подходов, включая удаление функциональных слов, использование таблицы частоты слов, удаление специальных слов и изменение формы слова.

*Удаление функциональных слов*

Функциональные слова включают предлоги, определители / местоимения, союзы, междометия и частицы дискурса. Удаление функциональных слов, которое часто может сделать запрос достаточно коротким, может быть использовано в качестве основы для формирования запроса в поисковых системах.

Некоторые слова не являются функциональными словами, но действуют как функциональные слова логически или структурно, например, «вид», «имя дизайнера» также может быть удалено из запроса:

Для длинных вопросов запрос, сделанный на этом шаге, все еще слишком длинный. Например, после удаления всех функциональных слов, о которых идет речь: «*Как называется редкое неврологическое заболевание с такими симптомами, как: непроизвольные движения (тики), ругань и бессвязные вокализации (хрюканье, крики и т.д.)?* ", мы получаем "*название редкого неврологического заболевания симптомы непроизвольных движений тики ругаются бессвязными вокализациями хрюкает криками"*. Длина этого запроса составляет 13 слов и недостаточно хороша, чтобы быть запросом для задач контроля качества.

*Использование таблицы частоты слов*

Другой способ сделать запрос короче — удалить часто используемые слова в запросе. Основная идея заключается в том, что чем чаще слово используется в языке, тем менее разборчиво это слово. Таким образом, AnswerBus реализует таблицу частоты слов. Для длинного запроса AnswerBus сортирует все слова в запросе и удаляет одно или несколько слов, которые определены как часто используемые в таблице частот.

*Изменение текстовой формы*

Некоторые слова в исходном вопросе преобразуются в другую форму, а затем помещаются в запрос. Обычно это глаголы, например, *окончание* Å *заканчивалось* вопросом «Когда закончился юрский период?» *У* Å возникает вопрос: «Сколько сердец у осьминога?»

# КАНДИДАТ ОТВЕТ EXTRACTION

На этом этапе AnswerBus загружает и обрабатывает документы, указанные в верхней части результатов поиска, возвращаемых различными поисковыми системами. Сначала он анализирует документы на предложения, а затем определяет, является ли предложение кандидатом на ответ, с помощью процесса сопоставления слов.

Инструмент сегментации предложений, используемый в AnswerBus, предназначен для обработки сложных веб-документов. Помимо удаления HTML-тегов, он исключает неконтекстовый контент; рассматривает некоторые специальные HTML-теги как обозначения границ предложения; и учитывает различные исключения форматирования.

Чтобы определить, является ли извлеченное предложение потенциально ответом на вопрос, AnswerBus классифицирует все слова в исходном вопросе или предложения в извлеченных документах на две категории: совпадающие слова и несоответствующие слова. Все слова, которые используются для формирования конкретного запроса поисковой системы, являются совпадающими словами. Остальные – несоответствующие слова.

Следующая формула используется для фильтрации извлеченных предложений.

*в*≥ *Q*−1+1

В этой формуле q — количество совпадающих слов в предложении; Q — общее количество совпадающих слов в запросе. Например, если запрос содержит три слова, то предложение-кандидат на ответ должно содержать не менее двух из них. Когда предложение соответствует условию, указанному в приведенной выше формуле, оно получит первичный балл, основанный на количестве совпадающих слов, которые оно содержит. В противном случае он получит оценку «0».

# РЕЙТИНГ ОТВЕТОВ

После извлечения предложений-кандидатов на ответ каждое предложение получило первичный балл. Предложения с оценкой «0» отбрасываются. Тем не менее, первичные оценки недостаточно надежны для суждения о том, является ли предложение реальным ответом. AnswerBus использует несколько методов для уточнения первичных оценок, включая определение типа вопроса, использование словаря для обеспечения качества, извлечение именованных сущностей, разрешение кореференции. Итоговая оценка, которая используется для определения ранга ответа, представляет собой комбинацию первичного балла и влияния всех различных факторов.

## Тип вопроса и словарь контроля качества

Почти все системы контроля качества используют тип вопроса для оценки ответа. Они классифицируют типы вопросов на основе типов ответов, ожидаемых пользователями. Например, «*Кто есть...?* " будет присвоен

Тип "ЛИЦО/ОРГАНИЗАЦИЯ"; в то время как «*Когда сделал... происходить?* " будет классифицирован как вопрос "ДАТА/ВРЕМЯ".

AnswerBus также использует тип вопроса в качестве важной информации, чтобы судить, может ли предложение быть ответом на вопрос. AnswerBus классифицирует вопросы на различные типы вопросов вместе с некоторыми параметрами. Например, AnswerBus классифицирует оба «*Как далеко...?* " и "*Насколько близко...?* " вопросы как вопрос DISTANCE, но он также дифференцирует эти два типа вопросов: единицу ответа на "*Как далеко...?* Скорее всего, это будет «миля», «километр», «световой год» и другие связанные с ними более крупные единицы, у него мало шансов быть «дюймом», «сантиметром» и т. Д. Для «*Насколько близко...?* " вопрос, единицей ответа на этот вопрос может быть любой из вышеперечисленных, в зависимости от контекста в вопросе, он даже может быть "нанометровым" или другим.

AnswerBus использует специальный словарь QA, базу данных, содержащую такого рода информацию о взаимосвязи слов между вопросами и ответами. Например, для записи слова «далеко» определение, приведенное в словаре, содержит «мили», «километры», «световые годы»; для слова «высокий» определение содержит «ноги», «метры». Словарь используется для различения типов вопросов и определения того, является ли предложение правильным ответом.

## Динамическое извлечение базовых именованных сущностей

Извлечение именованных сущностей широко используется в закрытых системах контроля качества для обеспечения поддержки выбора ответов. Это достигается путем предварительного пометки корпуса или базы знаний, на которых основана система контроля качества. Согласно [13], скорость маркировки составляет примерно 100 М/час. При условии, что, по одному вопросу, веб-система контроля качества должна обрабатывать 50 HTML-документов средней длины с общим размером 1 МЛН байт, время, которое необходимо увеличить для пометки этих документов, составит 36 секунд. Для такой системы существуют еще другие затраты времени, такие как сетевой ответ, передача по сети и другая обработка текста. Таким образом, обычные методы маркировки не являются неосуществимыми для систем контроля качества в режиме реального времени.

AnswerBus выполняет динамическое извлечение именованных сущностей при обработке предложений из веб-документов. AnswerBus извлекает только именованные сущности, соответствующие типам вопросов. Например, когда задается вопрос о том, «сколько денег», предложение с сущностью «ВАЛЮТА» получит более высокий ранг.

## Разрешение ссылок

Некоторые предложения содержат слова, такие как «он», «они», которые соотносятся с другими объектами, описанными в документе. Они не так хороши, как настоящая фраза существительного, чтобы быть ответами на вопросы, но они могут направлять пользователей к контекстным URL-ссылкам, которые содержат полные ответы. Таким образом, AnswerBus дает более низкие, но все же действительные оценки этим предложениям. Разрешение корреляции также является тяжелой задачей НЛП, такой как извлечение именованных сущностей. Вместо того, чтобы использовать полное разрешение coreference, AnswerBus решает только основные ссылки в соседних предложениях. При обнаружении этого типа корреляции более позднее предложение получает часть оценки от своего предыдущего предложения.

## Позиция попадания и уверенность в поисковой системе

Рейтинг предложения также может быть связан с позицией, в которой его исходный документ находится в списке попаданий, возвращаемом поисковой системой. Предложение, извлеченное из первого удара, получает самый высокий балл, связанный с позициями удара, и счет уменьшается по мере продвижения позиции вниз. Документы, возвращенные разными поисковыми системами, также могут получать разные оценки.

## Избыточность

Разные поисковые системы могут получить один и тот же документ для вопроса. Одинаковые или очень похожие предложения могут быть извлечены из одного и того же документа или из разных документов. Это приводит к избыточности предложения кандидата на ответ.

Чтобы обнаружить избыточность предложений, AnswerBus сравнивает до сих пор высоко оцененные предложения друг с другом. Однако знаки препинания, пробелы, другие специальные символы и слова с очень высокой частотой не рассматриваются для того, чтобы проводить сравнения между предложениями.

# ОЦЕНКА

200 вопросов TREC 8 были использованы для оценки производительности ответов на вопросы AnswerBus, а также для сравнения его производительности с производительностью четырех других подобных систем. Была написана компьютерная программа для отправки вопросов один за другим в AnswerBus, START[11], LCC, IONAUT и QuASM; а также получить ответы из систем. После получения ответа на вопрос временной интервал

До отправки следующего вопроса было предоставлено 10 секунд, чтобы гарантировать, что работа системы над предыдущим не повлияет на производительность системы. Время ответа на каждый вопрос системами и длина возвращенных ответов были записаны. Чтобы свести к минимуму влияние производительности сети на изменения времени отклика систем, ответы из этих систем были получены одновременно с использованием одного и того же компьютера.

За исключением IONAUT, пять лучших ответов на каждый вопрос из систем затем оценивались вручную, чтобы определить точность ответов. ИОНАВТ возвращал длинные отрывки информации в качестве ответов на вопросы. Оценка информации выходила за рамки ресурсов этого проекта. Таким образом, для остальных систем ответы сначала сравнивались с ключами ответов, предоставленными TREC. Поскольку эти системы основаны на Интернете, который отличается от большого корпуса новостных лент, на которых основаны вопросы TREC, некоторые ответы, которые отличались от ключей ответов, также были оценены как правильные. Ответы также были тщательно изучены, чтобы убедиться, что контекстуально правильные ответы находятся.

В таблице 1 представлена производительность AnswerBus и других четырех систем. Он предоставляет количество правильных ответов в первой пятерке и первом ответе системы, стандартные баллы NIST, максимальное, минимальное и среднее время отклика, измеряемое в секундах, стандартное отклонение времени отклика и среднюю длину возвращенных ответов.

Из таблицы 1 видно, что AnswerBus превосходит другие аналогичные системы как по точности, так и по времени отклика. AnswerBus также вернул более краткие ответы, чем другие системы.

Таблица 1 Производительность онлайновых систем ответов на вопросы

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Системы** | **Правильный ТОП 5** | **Правильный ТОП 1** | **НИСТ**  **Счёт** | ***Тмакс***  ***(с)*** | ***Тмин***  ***(с)*** | ***Тмин***  ***(с)*** | ***Tstd dev*** | ***Лмеан***  ***(байт)*** |
| **ОтветБус** | 141 | 120 | 64.18% | 15.06 | 3.79 | 7.20 | 3.07 | 141 |
| **ИОНАВТ** |  |  |  | 44.88 | 2.78 | 12.51 | 6.81 | 1312 |
| **ЛКЦ** | 97 | 75 | 41.73% | 342.52 | 4.30 | 44.24 | 32.63 | 178 |
| **КуАСМ** | 13 | 7 | 4.45% | 284.29 | 2.61 | 20.72 | 33.92 | 1766 |
| **НАЧАЛО** | 29 | 29 | 14.50% | 62.07 | 2.02 | 9.84 | 7.45 |  |

Таблица 2 Производительность AnswerBus по сравнению с Малдером

|  |  |  |  |
| --- | --- | --- | --- |
| **Системы** | **Правильные топ-5 ответов** | **Правильный топ-1 ответов** | **Оценка** NIST |
| **ОтветБус** | 70.5% | 60% | 64.2% |
| **Малдер** |  | 34% |  |

Примечание  : Источник [12].

Аналогичные данные для двух других известных систем контроля качества, Малдера и Вебклопедии, не были получены из-за их автономного статуса, когда проводился эксперимент. По данным [12], Малдер правильно ответил на 34% вопросов TREC-8. В таблице 2 приведено некоторое сравнение между AnswerBus и Малдером.

# Ссылки

1. Стивен Эбни, Майкл Коллинз и Амит Сингхал. Извлечение ответа. *Материалы АНЛП 2000 года.*  Сиэтл, штат Вашингтон. 29 апреля - 3 мая 2000 года.
2. Юджин Агихтей, Стив Лоуренс, Луис Гравано. Изучение преобразований запросов для поисковых систем для ответов на вопросы*.*  *Десятая Всемирная веб-конференция*. 1-5 мая 2001 года, Гонконг, Китай.
3. Эрик Брек, Джон Бургер, Лиза Ферро, Дэвид Хаус, Марк Лайт и Индержит Мани. Система с именем Qanda. *Восьмая текстовая конференция REtrieval (TREC-8)*. Гейтерсберг, MD. 17-19 ноября 1999 года.
4. Джей Будзик и Кристиан Дж. Обучение для ответов на вопросы и классификации текста: интеграция основанных на знаниях и статистических методов*.*  *Семинар* *AAAI по тексту*

*Классификация.* Менло-Парк, Калифорния, 1998

1. Питер Кларк, Джон Томпсон и Брюс Портер. Основанный на знаниях подход к ответам на вопросы. *Осенний симпозиум* *AAAI'99 по системам вопросов и ответов*. Орландо, Флорида. 1999 год.
2. E. J. Glover, S. Lawrence, M. D. Gordon, W. P. Birmingham, and C. L. Giles, "Web Search -- Your Way", *Communications of the ACM*, To appear.

1. 6DQGD+DUDEDJLX'DQ0ROGRYDQ0DULXV3DúFD Михай Сурденану, Рада Михалча, Роксана Гирджу,

Василе Рус, Финли Лакатусу, Поль Морареску и

Разван Бунеску. Комплекс ответов, список и

Контекстные вопросы с сервером ответов на вопросы LCC. *Десятая текстовая конференция (TREC-10).* Гейтерсберг, MD. 13-16 ноября 2001 года.

1. Санда Харабаджиу, Мариус Паска и Стивен Майорано. *Эксперименты с текстовыми ответами на вопросы с открытым доменом*. *КОЛИНГ-2000. Ассоциация компьютерной лингвистики/Морган Кауфманн*, август 2000 года.
2. Линетт Хиршман и Роберт Гайзаускас. Ответ на вопрос на естественном языке: взгляд отсюда. *Инженерия естественного языка,* 2001.
3. Эдуард Хови, Лори Гербер, Ульф Хермякоб, Майкл Джанк и Чин-Тис Лин. Ответы на вопросы в Вебклопедии. Девятая текстовая конференция (TREC-9). Гейтерсберг, MD. 13-16 ноября 2000 года.
4. Борис Кац, От обработки предложений к доступу к информации во Всемирной паутине.

*Весенний симпозиум AAAI по обработке естественного языка для Всемирной паутины.* Стэнфорд, Калифорния. 1997 год.

1. Коди К. Т. Квок, Орен Этциони и Дэниел С. Уэлд. Масштабирование ответов на вопросы в Интернете*.*

*Десятая Всемирная веб-конференция*. Гонконг, Китай. 1-5 мая 2001 года.

1. Рохини Шрихари и Вэй Ли. Извлечение информации Поддерживает ответы на вопросы. *Восьмая текстовая конференция REtrieval (TREC-8).* Гейтерсберг, MD. 17-19 ноября 1999 года.

1. http://www.answerbus.com/ [↑](#footnote-ref-1)
2. http://www2.languagecomputer.com/qa/head.html [↑](#footnote-ref-2)
3. http://ciir.cs.umass.edu/~reu2/ [↑](#footnote-ref-3)
4. http://www.ionaut.com:8400/ [↑](#footnote-ref-4)
5. http://www.ai.mit.edu/projects/infolab/ [↑](#footnote-ref-5)
6. http://www.isi.edu/natural-language/projects/webclopedia

   /исследование.html#ВОПРОС [↑](#footnote-ref-6)
7. http://babel.altavista.com/ [↑](#footnote-ref-7)