Система вопросно-ответного поиск Lasso Q/A System по материалам конференции TREC-1999

Доклад Устинова В. Д.

Научные руководители:
Большакова Е. И.
Добров Б. В.

2008

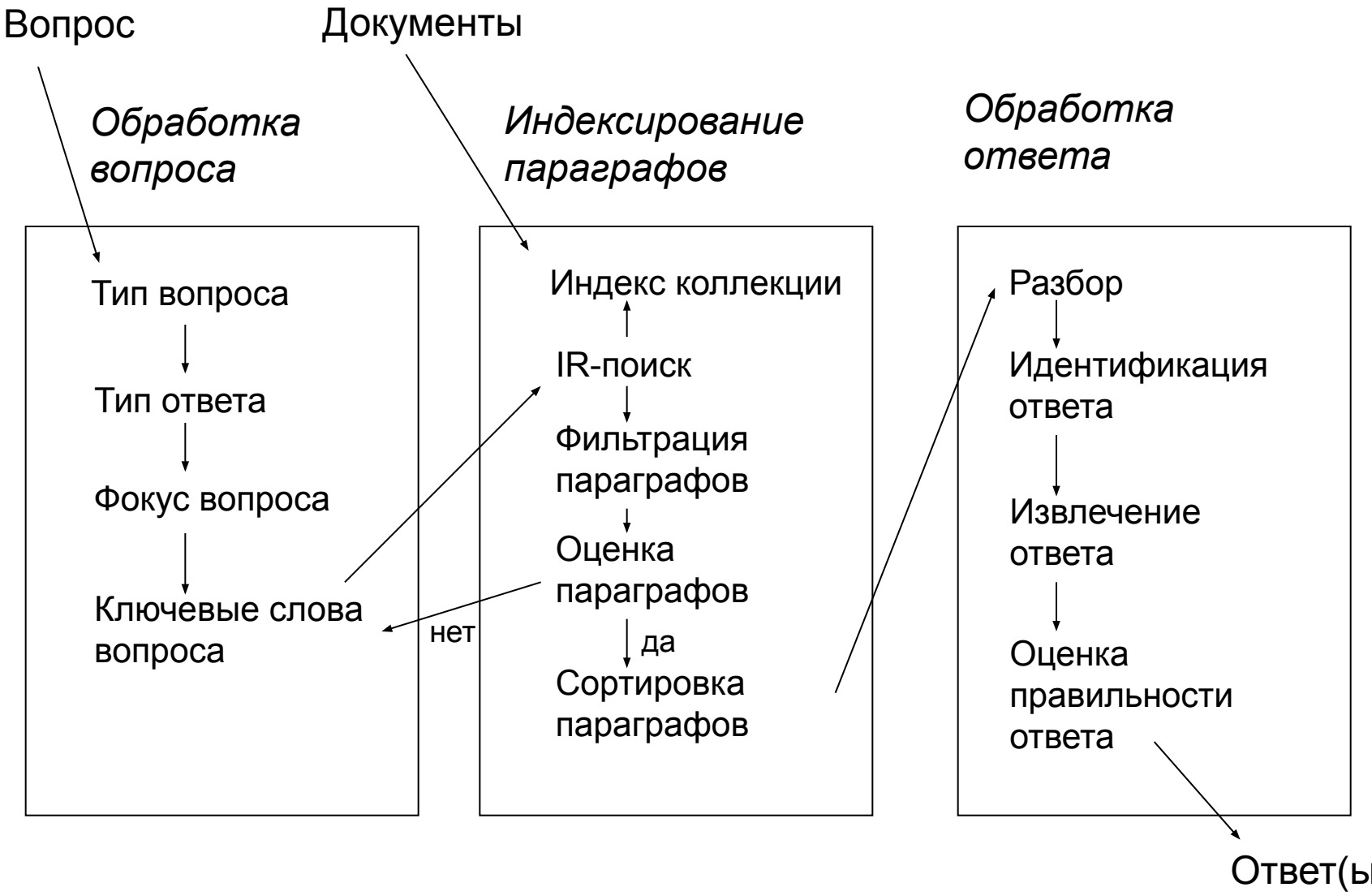
Примеры вопросов

- В каком году родился Пушкин?
- Кто был первым космонавтом?
- Сколько Microsoft потратила на исследования в 2004 году?
- Какое расстояние между Москвой и Питером?
- Где находится Тадж Махал?
- Назовите фильм, получивший Оскар.
- Почему Гугл купил компанию «Бегун»?
- Кого Зенит победил на Чемпионате Европы в 2008?

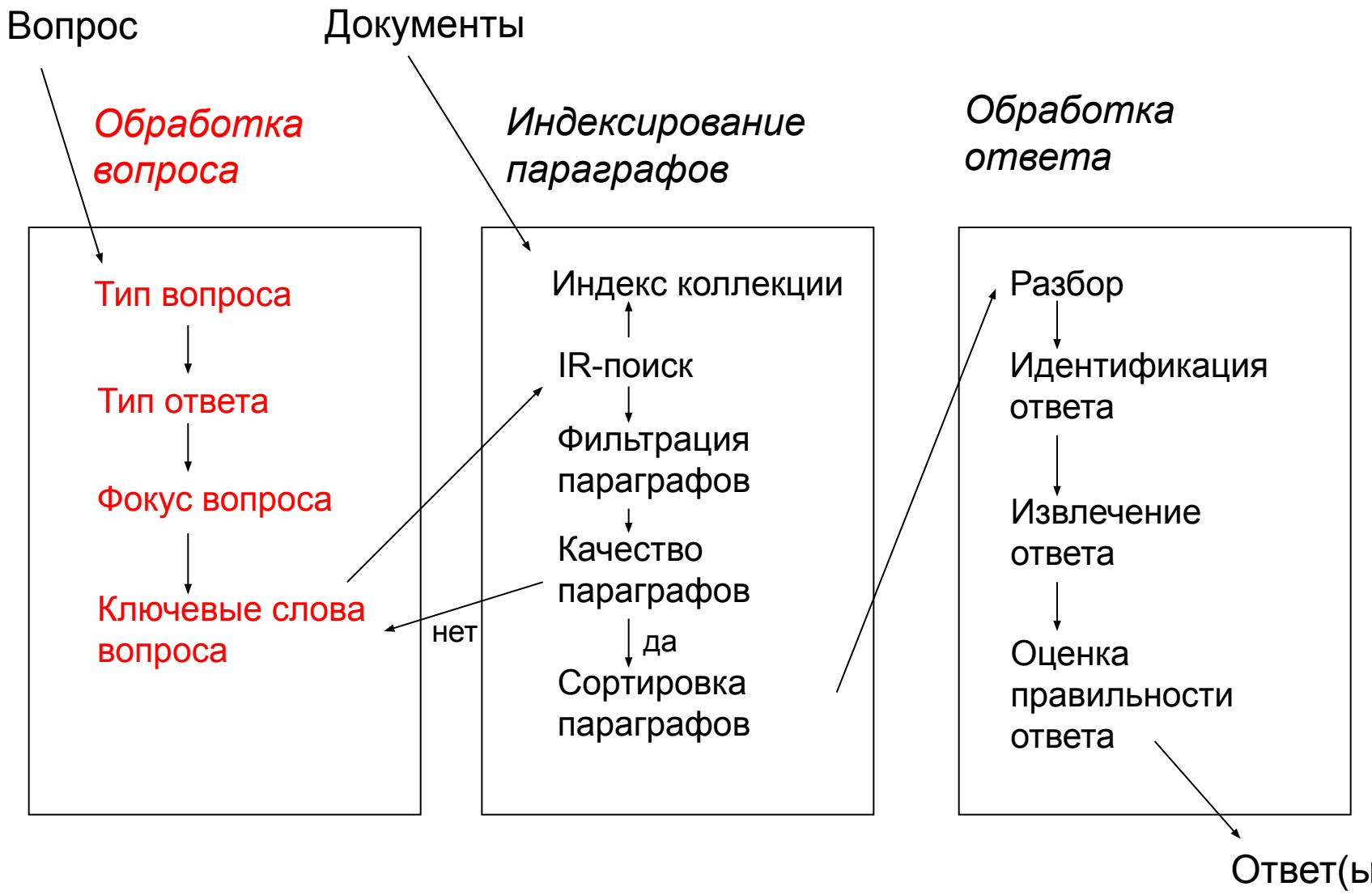
Техники вопросно-ответного поиска

- Information Retrieval
  - находит документ, который может содержать ответ на заданный вопрос
- Information Extraction
  - строит большую базу знаний и выдает четкий ответ, подставляя объекты в некоторый шаблон ответа
- Комбинация?

Архитектура Lasso Q/A System



Архитектура Lasso Q/A System



Пример обработки вопроса

Класс	Подкласс	Тип ответа	Пример вопроса	Фокус
what	basic what	money / number / definition / title / nnp / undefined	What was the monetary value of monetary value the Nobel Peace Prize in 1989?	monetary value
	what-who	person / organisation	What costume designer decided costume designer organization that Michael Jackson should only wear one glove?	costume desginer
	what-when	date	In what year did Ireland elect its first woman president?	year
	what-where	location	What is the capital of Uruguay?	capital
who		person / organisation	Who is the author ot the book "The Iron Lady: A Biography of Margaret Thatcher"?	author

Правила определения ключевых слов

1. Если в вопросе есть цитата с кавычками, все значимые слова (не стоп-слова) цитаты добавляются к списку ключевых слов.
2. Все имена собственные.
3. Все нарицательные имена с прилагательными.
4. Все остальные нарицательные имена
5. Все существительные с прилагательными
6. Все остальные существительные

7. Все глаголы
8. Фокус вопроса

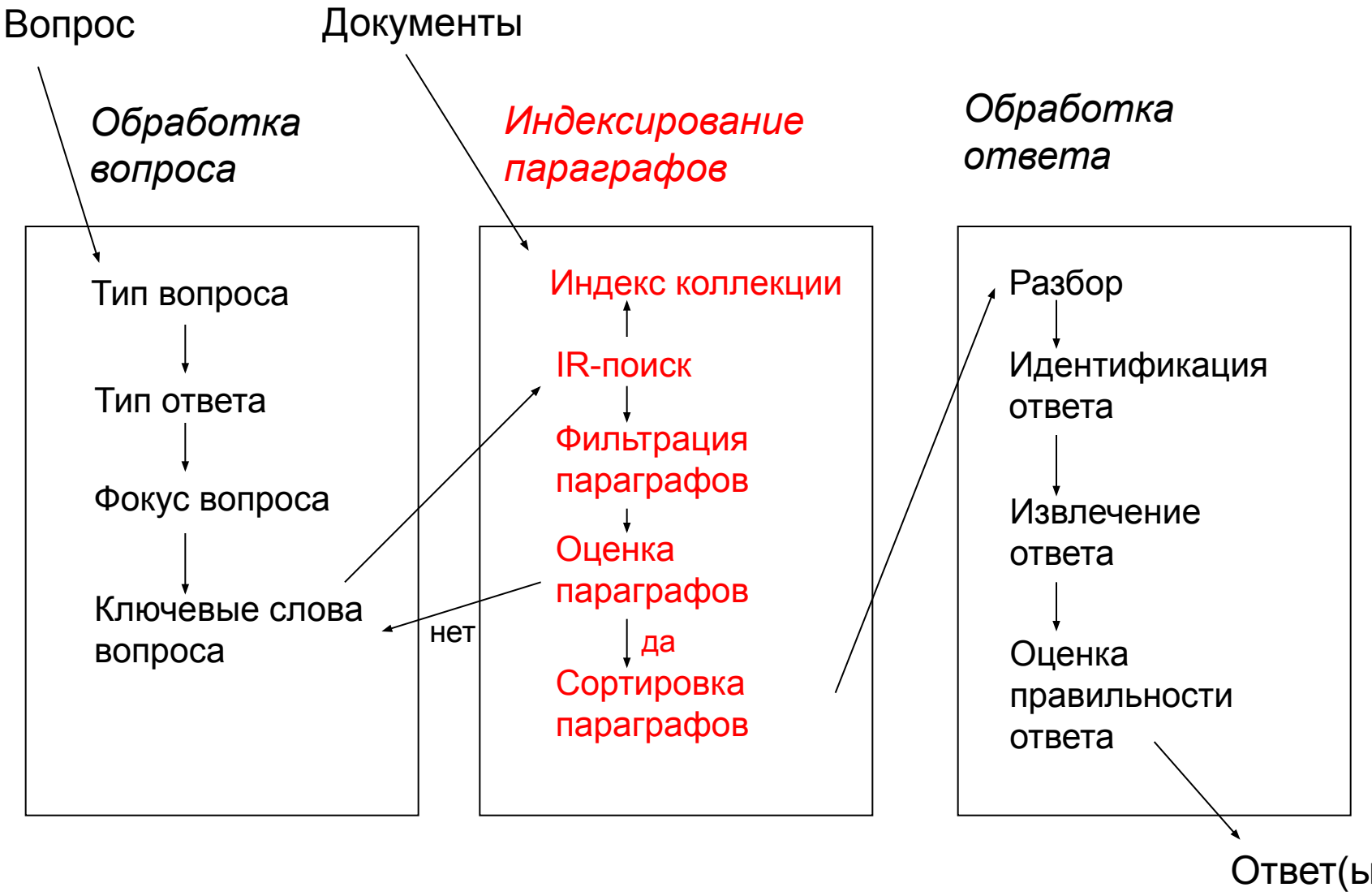
Примеры ключевых слов

- What is the name of the «female» counterpart to El Nino, which results in cooling temperatures and very dry weather ?

- female El Nino dry weather cooling temperatures
- female El Nino dry weather cooling
- female El Nino dry weather
- female El Nino dry
- female El Nino
- female El

- How much could you rent a Volkswagen bug for in 1966 ?
- Volkswagen bug
- Volkswagen bug rent

Архитектура Lasso Q/A System



IR-поиск

Построение индекса коллекции:

1. Нормализация SGML-тегов
2. Исключение лишних символов
3. Разделение на слова
4. Нормализация (стемминг) слов
5. Расчет локальных и глобальных весов
6. Построение общего словаря коллекции
7. Создание инвертированного индексного файла

Особенности IR-поиска:

1. Булевское индексирование вместо Векторного

Фильтрация параграфов

Оператор PARAGRAPH n – действует как AND, но только в пределах n параграфов, а не в пределах всего документа

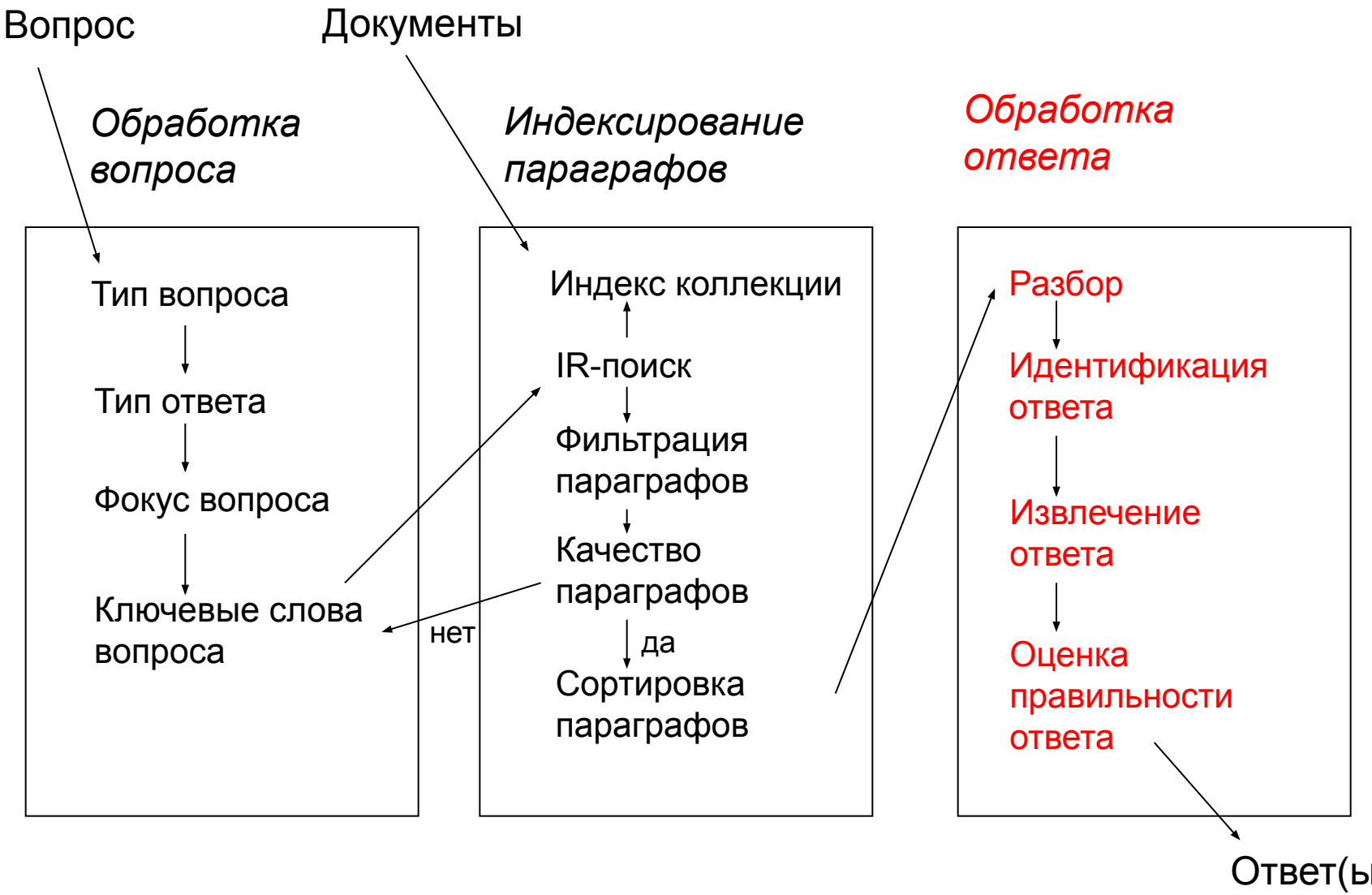
Деление на предложения – с помощью знаков препинания

Деление на параграфы – с помощью HTML-тегов, пустых строк, отступов.

Оценка и сортировка параграфов

- Оценивается не просто параграф, а окно внутри параграфа
- Оценка окна =
  + max (Оценка слов в том же порядке)
  – max (Расстояние между самыми дальними ключевыми словами)
  – min (Количество недостающих ключевых слов)

Архитектура Lasso Q/A System



Разбор (парсер) + идентификация ответов

1. Определение частей речи
2. Выделение именованных сущностей с помощью словарей Gazetteers и WorldNet.
3. Распознавание имен персон, названий организаций, мест валют и продуктов, дат с помощью эвристических правил.

Все те же возможности наиболее широко используются в системах Information Extraction

=> Все слова, подходящие по типу под тип ответа, помечаются как ответы-кандидаты.

Извлечение ответа и оценка

Оценивается ответ-окно. Оценка ответа-окна считается так:

+32 * Оценка за совпадения с ключевыми словами

+16 * Оценка за пунктуацию (за ответом следует знак препинания)

+16 * Оценка за слова из вопроса, следующие сразу за ответом-кандидатом после запятой

+16 * Оценка слова из вопроса, найденные в том же поддереве разбора

+16 * Оценка за слова из вопроса в том же предложении

+16 * Оценка за общее количество ключевых слов во всем ответе-окне

-4 * sqrt (Оценка за расстояние между ответом-кандидатом и другими словами вопроса в ответе-окне)

Примеры вопросов и ответов

Вопрос: What is the name of the rare neurological disease with symptoms such as : involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc)?

Ответ (короткий), оценка: 284.40 - who said she has both Tourette's Syndrome and

Вопрос: Where is the actress Marion Davies, buried ?

Ответ (короткий), оценка: 142.56 - from the fountain inside Hollywood Cemetery

Вопрос: Where is the Taj Mahal ?

Ответ (длинный), оценка: 408.00 - list of more than 360 cities throughout the world includes the Great Reef in Australia, the Taj Mahal in India, Chartre's Cathedral in France, and Seregenti National Park in Tanzania. The four sites Japan has listed include

Вопрос: What is the nationality of Pope John Paul II ?

Ответ (длинный), оценка: 407.06 - stabilize the country with its (long) help, the Catholic hierarchy stoutly held out for pluralism, in large part at the urging of Polish-born Pope John Paul II. When the Pope emphatically defended the Solidarity trade union during a 1987 tour of the