

Основы обработки текстов

Лекция 11

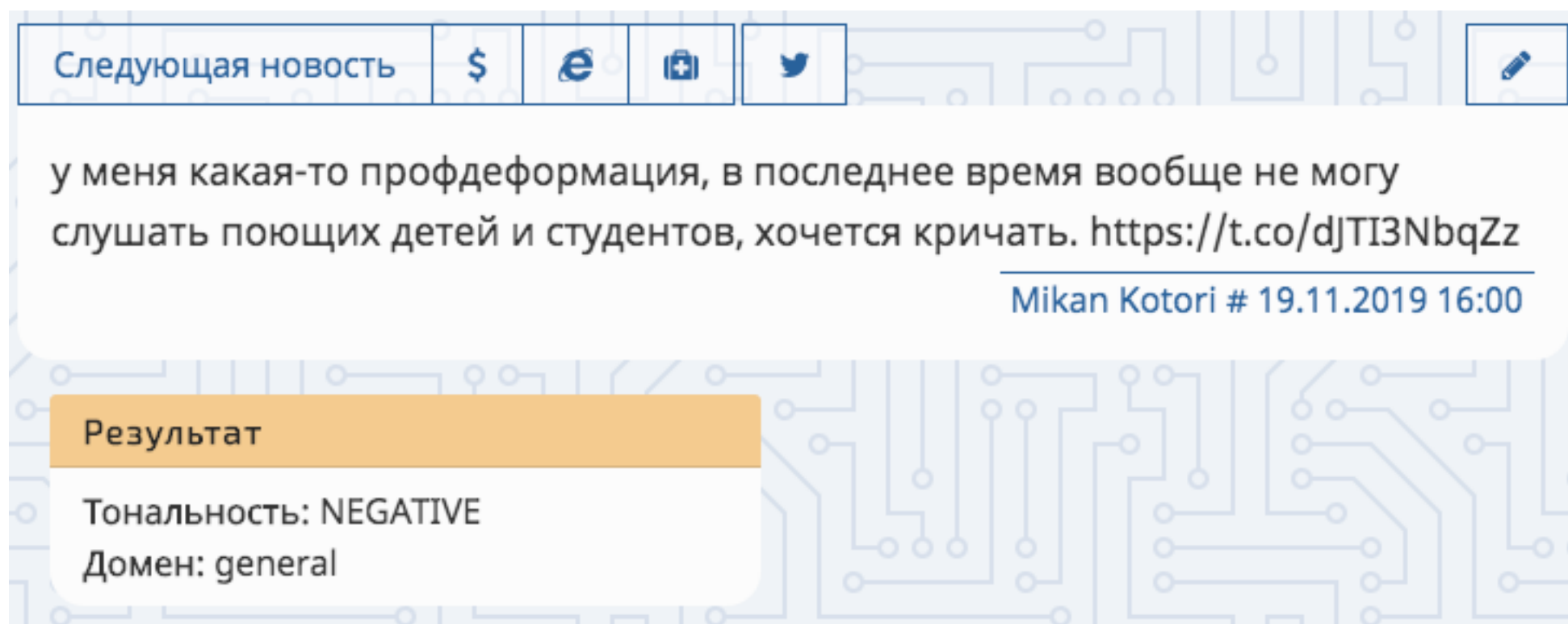
Прикладные задачи обработки текстов

План

- Анализ тональности текстов
- Вопросно-ответные системы
- Автоматическое реферирование

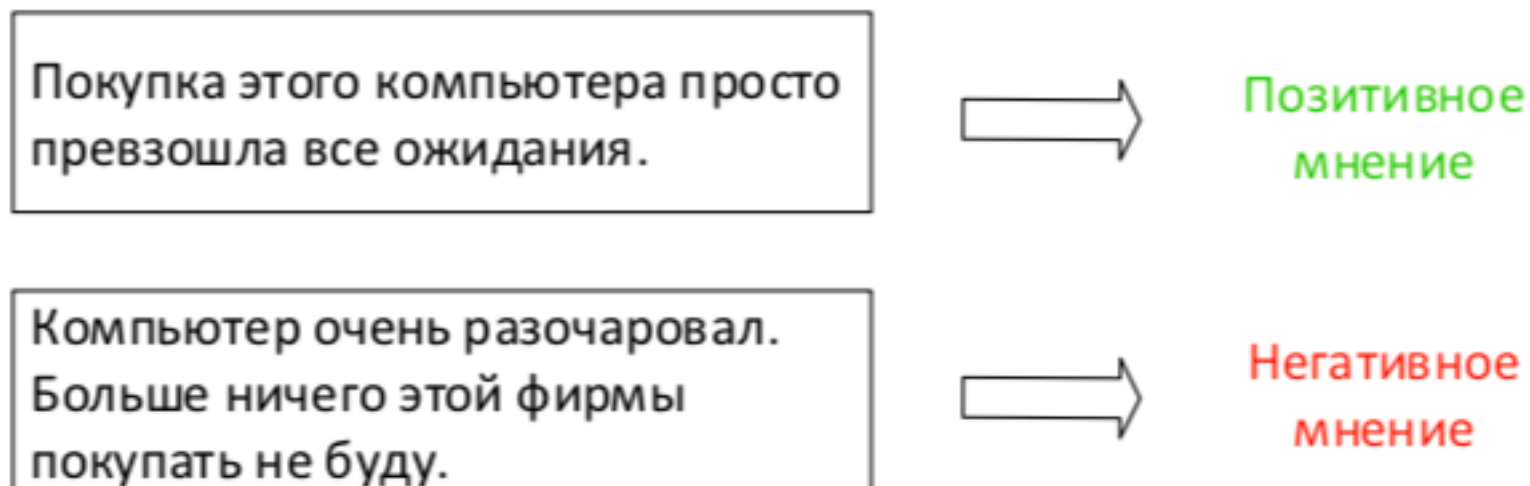
Анализ тональности текстов

- Sentiment analysis, Opinion mining
- **Тональность** текста - эмоциональное отношение автора к некоторому объекту



Texterra: определение тональности

Применение



- Маркетинговые исследования и репетиционный менеджмент
- Рекомендательные системы
- Анализ финансовых рынков
- Политологические исследования
- Социологические исследования и др.

Формальная постановка

- Мнение $o(d) = (e, a, c, h, t) \in O$
 - e - объект, по отношению к которому выражается мнение
 - a - свойство / атрибут / аспект объекта
 - c - тональность мнения
 - h - автор мнения
 - t - время выражения мнения
- Для заданного корпуса текстовых D документов построить функцию F :

$$F : D \rightarrow O$$

Тональность мнения

- Наиболее частая постановка: позитивная, негативная, нейтральная
 - Трех-классовую классификацию иногда разделяют на две бинарные
 - Субъективность / объективность
 - Полярность (+ / -)
- Альтернатива: шкала (например, от -10 до 10)
- Альтернатива: эмоции

Подходы к решению задачи

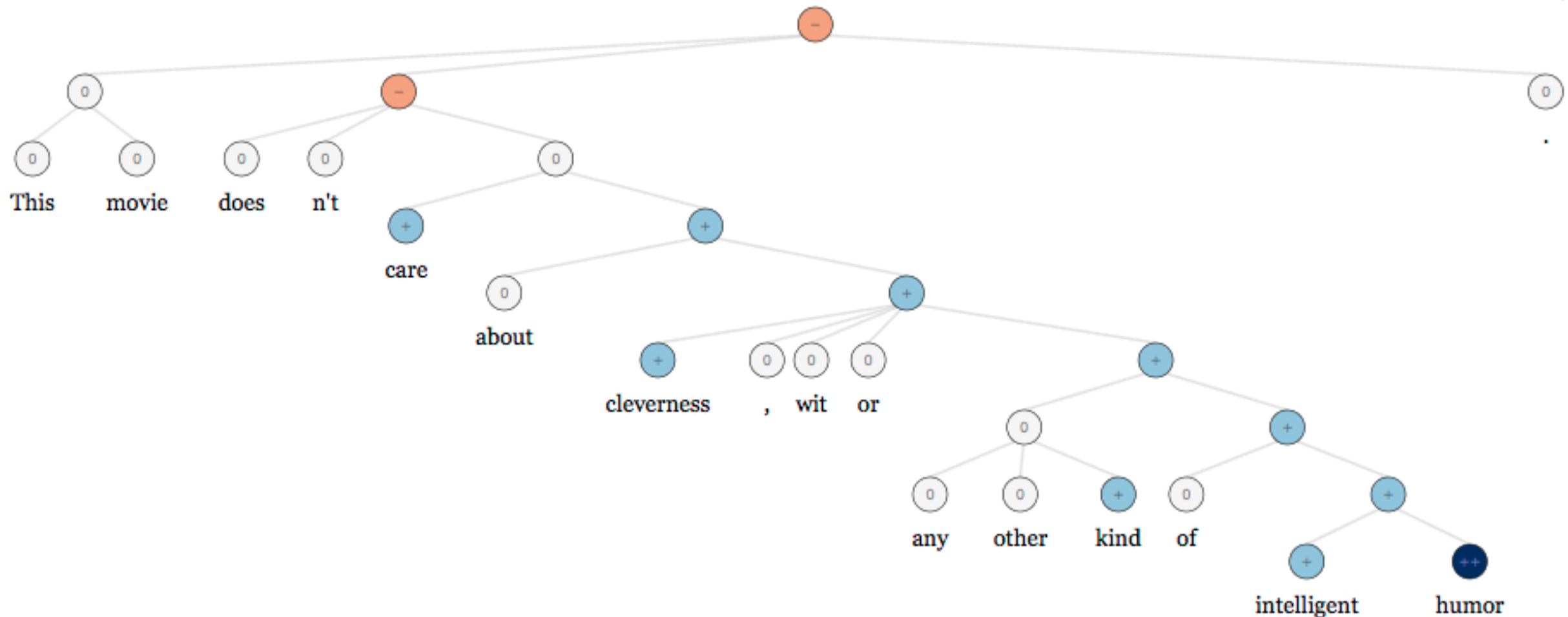
- Подход основан на словарях
 - Основная проблема - многозначность (*старый* друг, *старый* матрац)
- Машинное обучение с учителем
 - Любые классификаторы
 - Мешок слов, word2vec
- Использование специализированных тезаурусов

Семантические тезаурусы с эмоциональной составляющей

- Wordnet-Affect
- SentiWordNet
- SenticNet

# POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	1740	0.125	0	able#1	(usually followed by `to') having the necessary means or skill or know-how or authority to do something; "able to swim"; "she was able to program her computer"; "we were at last able to buy a car"; "able to get a grant for the project"
a	2098	0	0.75	unable#1	(usually followed by `to') not having the necessary means or skill or know-how; "unable to get to town without a car"; "unable to obtain funds"

Корпус синтаксических деревьев с эмоциональной разметкой



nlp.stanford.edu/sentiment/treebank.html

Аспектно-ориентированный анализ

Были в этом ресторане в апреле месяце. Остались хорошие впечатления. Началось с того, что пришли туда в субботу, зал набит, мест не было. Расстроились. Стояли перед баром с подругой, ждали общего друга, чтобы решить, куда пойти-куда податься. Ждали минут 10 и только он пришел, как к нам вышла хостесс и вежливо предложила забронированный столик, объяснив, что мы можем там посидеть пока не появятся "хозяева" (к нашей радости, они так и не объявились).

```
<aspect category="Whole" from="12" mark="Rel" sentiment="neutral" term="ресторане" to="21" type="explicit"/>
```

```
<aspect category="Interior" from="113" mark="Rel" sentiment="negative" term="зал набит" to="122" type="fct"/>
```

```
<aspect category="Interior" from="124" mark="Rel" sentiment="negative" term="мест не было" to="136" type="fct"/>
```

```
<aspect category="Service" from="293" mark="Rel" sentiment="positive" term="хостесс" to="300" type="explicit"/>
```

```
<aspect category="Service" from="303" mark="Rel" sentiment="positive" term="вежливо" to="310" type="implicit"/>
```

```
<aspect category="Service" from="311" mark="Rel" sentiment="positive" term="предложила забронированный столик" to="344" type="fct"/>
```

Выявление аспектных терминов

- Классификация последовательности с разметкой BIO и др. (см. Лекцию 2)

Оценка качества

- Точность / полнота
- Отдельно для каждой подзадачи, например для SentiRuEval 2014
 - Task A: automatic extraction of explicit aspects
 - Task B: automatic extraction of all aspects including sentiment facts
 - Task C: extraction of sentiments towards explicit aspects
 - Task D: automatic categorization of explicit aspects into aspect categories
 - Task E: sentiment analysis of the whole review on aspect categories.

Альтернативные постановки

- Сравнительные анализ (comparative sentence mining)

Объективы у Canon лучше, чем у Nikon

- Подзадачи:
 - Поиск сравнительных предложений
 - Извлечение отношений

<сравнительное слово>, <Аспект>, <Сущность 1>, <Сущность 2>
<лучше>, <оптика>, <Canon>, <Nikon>

Альтернативные постановки

- Предсказание эмоций с точки зрения читателей

$$o(d) = (e, a, c, h, r, t) \in O$$

- r - читатель
- Нейтральный по стилю новостной текст может вызвать сильно эмоциональную реакцию у читателя

Нападающий туринского «Ювентуса» Криштиану Роналду не смог забить гол, находясь в метре от пустых ворот. Футболист умудрился заблокировать собственный удар.

Альтернативные постановки

- Обнаружение сарказма и иронии

Мне нравятся эти пятна горчицы на твоей толстовке.
Они подчеркивают цвет твоих глаз.

Я НЕ ХОТЕЛА ВАС ОБИДЕТЬ,
СЛУЧАЙНО ПРОСТО ПОВЕЗЛО!



Вопросно-ответные системы

Какой национальности бывший папа римский Бенедикт XVI?

Ватикан выступил во вторник, 12 мая, с опровержением информации о том, что Папа Римский Бенедикт XVI в юности состоял в гитлерюгенде. "Йозеф Рацингер (имя понтифика, **немца по национальности**) никогда не состоял в гитлерюгенде - идеологической нацистской организации.

Короткий фрагмент текста, не URL

Ответ: **Немец**

Примеры систем

 **WolframAlpha**™ computational knowledge engine

Enter what you want to calculate or know about:

How far is San Francisco from Moscow?

[Examples](#) [Random](#)

Assuming Moscow (Russia) | Use [Moscow \(Idaho, USA\)](#) or [more](#) instead

Input interpretation:
Moscow to San Francisco, California, United States

Distance: [Show non-metric units](#)

9472 km (kilometers)

Direct travel times: [More](#)

aircraft (550 mph)	10 hours 40 minutes
sound	7 hours 40 minutes
light in fiber	44.3 ms (milliseconds)
light in vacuum	31.6 ms (milliseconds)

(assuming constant-speed great-circle path)

AT&T 7:36 AM

“Today do I need an umbrella Ella Ella a a a a”

Yes, it's likely to rain today:

57° H: 57° L: 36°

8:00 AM 70% 57°

9:00 AM 70% 57°

10:00 AM 80% 55°



Типы вопросов

О фактах

Какая обычная высота жирафа?
Где расположен главный офис Google ?

Списки

Какие страны экспортируют нефть?
Какие названия имеют штаты США?

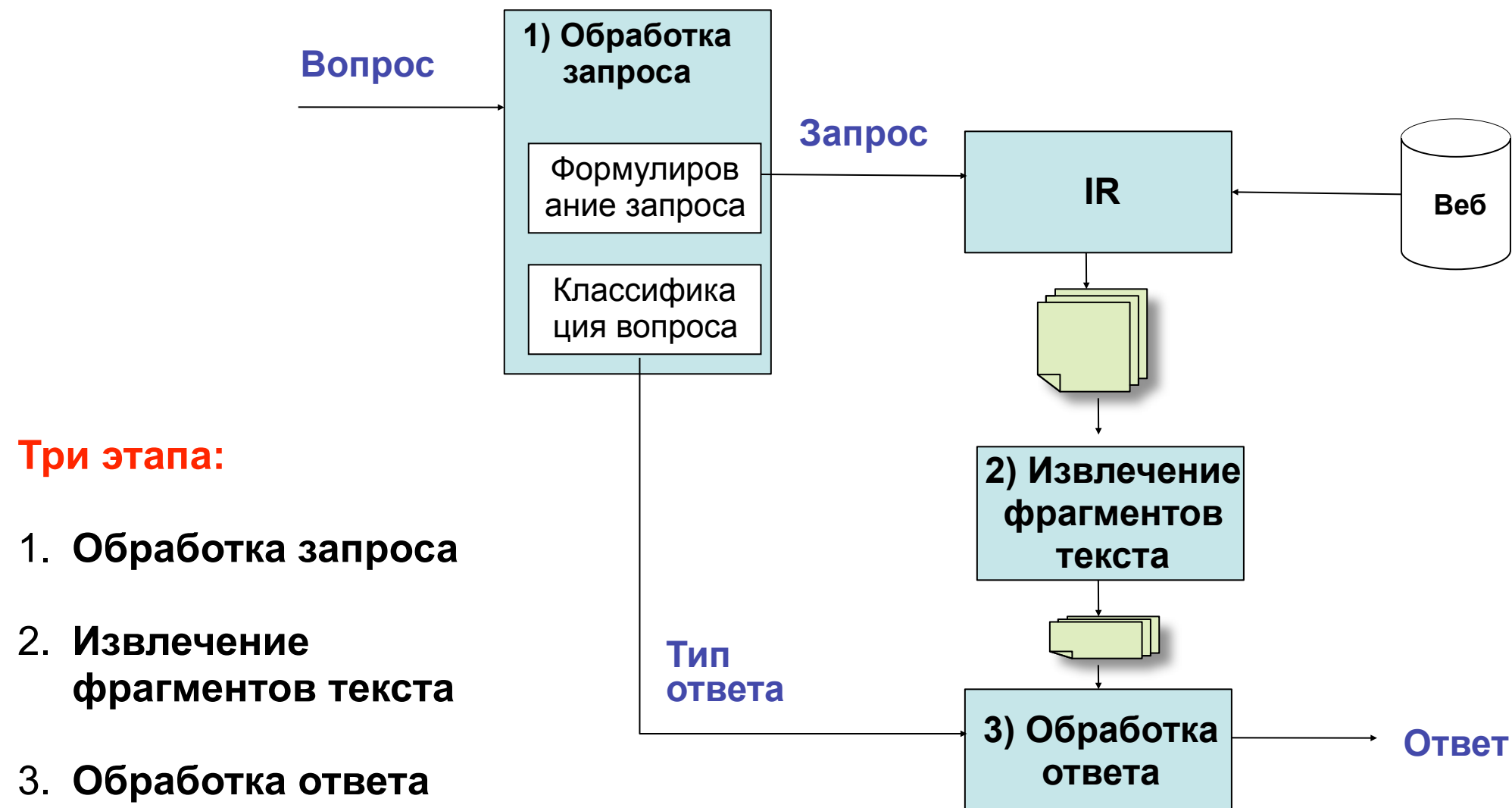
Определения

Кто такой Франсуа Томбалбай?
Что такое квазар?

Вопросы о фактах

- Ответом служит простой факт
 - Примеры:
 - Где расположен Лувр?
 - Какая называется валюта Китая?
 - Какой официальный язык Алжира?
- Существует большая разница между постановкой вопроса и описанием ответа в тексте
 - Какая компания является лидером по производству открыток?
 - Компания "Арт и Дизайн" десять лет назад создала в России практически новый рынок. Теперь она является лидером среди отечественных производителей поздравительных открыток.

Типичная архитектура QA-систем



Обработка запроса

- Из вопроса на естественном языке извлекаем:
 - ключевые слова для запроса к информационно-поисковой системе
(Формулирование запроса)
 - Тип ответа, специфицирующий класс сущности, возвращаемой в качестве ответа
(Классификация вопроса)

Формулирование запроса

- Извлечь ключевые термины из вопроса
– возможно расширить вопрос лексически/
семантически близкими словами
- Вопрос моделируется как множество
КЛЮЧЕВЫХ СЛОВ

Question (from TREC QA track)	Lexical terms
Q002: What was the monetary value of the Nobel Peace Prize in 1989?	monetary, value, Nobel, Peace, Prize, 1989
Q003: What does the Peugeot company manufacture?	Peugeot, company, manufacture
Q004: How much did Mercury spend on advertising in 1993?	Mercury, spend, advertising, 1993
Q005: What is the name of the managing director of Apricot Computer?	name, managing, director, Apricot, Computer

Формулирование запроса

- Применение правил для переформулирования вопроса
 - к форме подстроки декларативного ответа
 - “когда был придуман лазер” → “лазер был придуман”
 - Послать переформулированный запрос информационно-поисковой системе
 - Правила (Lin 07)
 - wh-word did A verb B → A verb-ed B
 - Where is A → A is located in

Классификация вопросов

- Классификация вопросов по ожидаемому ответу

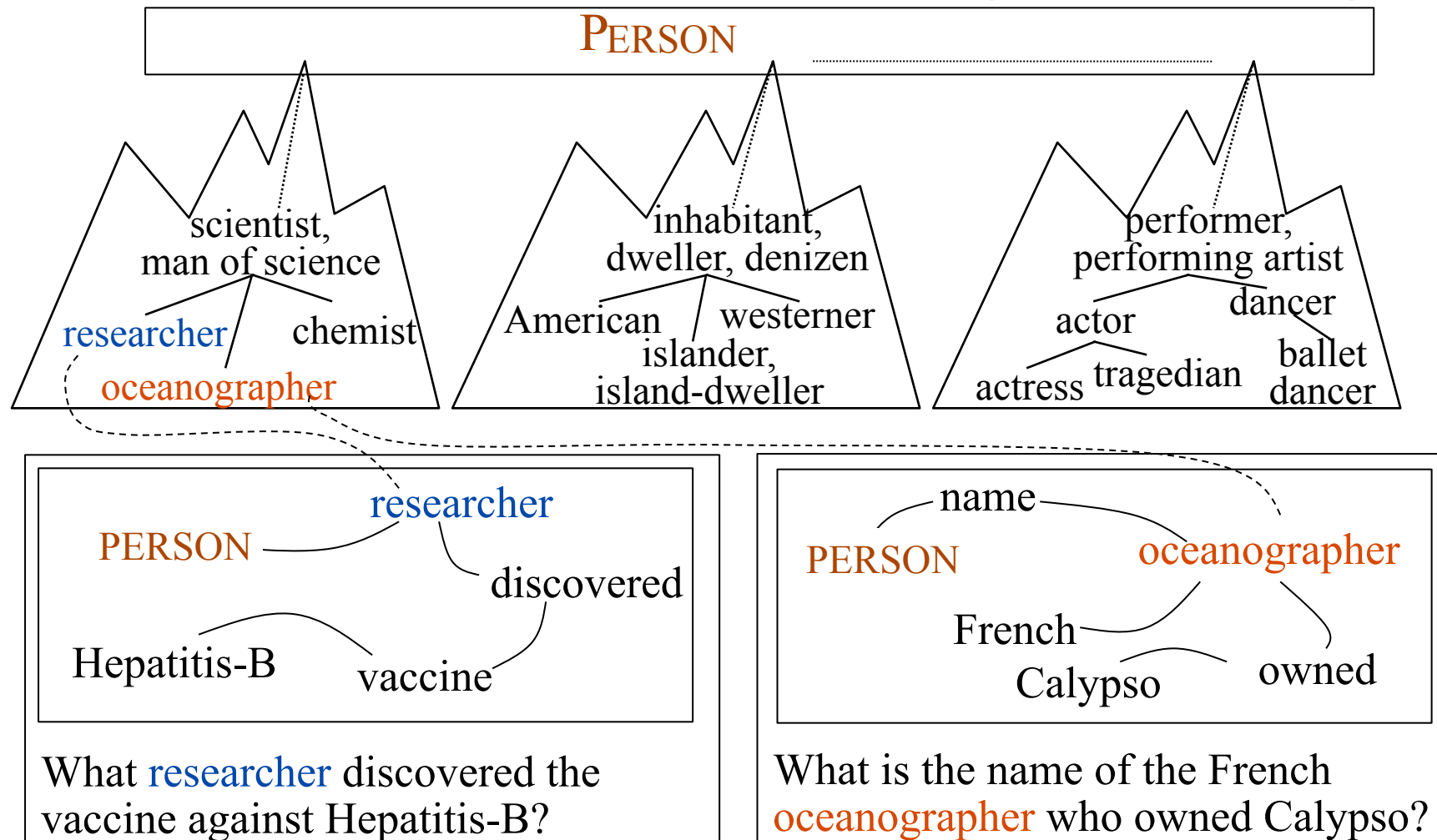
Вопрос	Основа вопроса	Тип ответа
Q555: What was the name of Titanic's captain?	What	Person
Q654: What U.S. Government agency registers trademarks?	What	Organization
Q162: What is the capital of Kosovo?	What	City
Q661: How much does one ton of cement cost?	How much	Quantity

Определение типа ответа

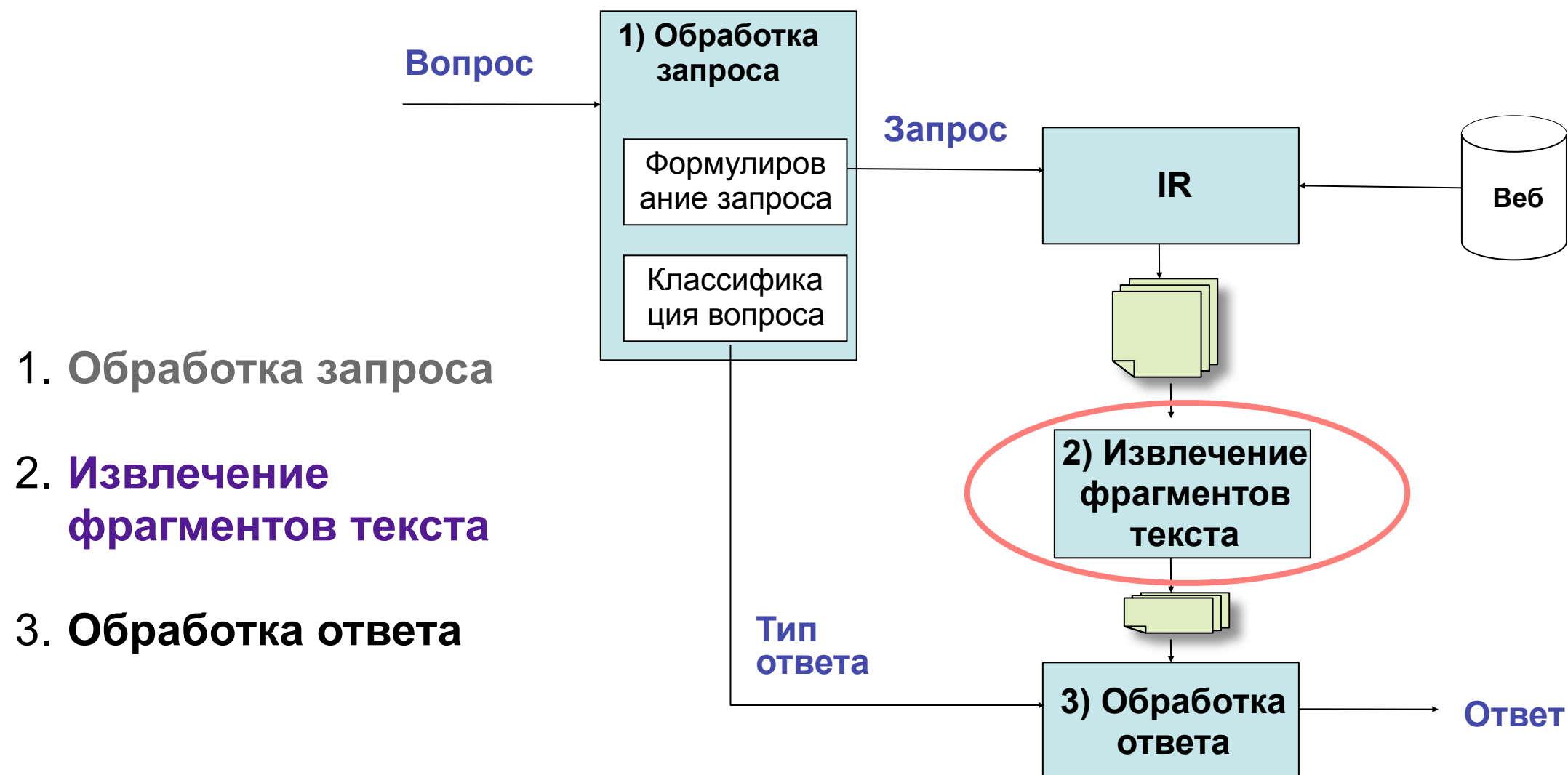
- В некоторых случаях тип ответа можно определить по вопросу
 - Почему → Причина
 - Когда → Дата
- Для многозначных вопросов использовать дополнительные понятия в вопросе
 - **What** was the name of Titanic's **captain**?
 - **What** U.S. Government **agency** registers trademarks?
 - **What** is the **capital** of Kosovo?
- Машинное обучение (если есть размеченный корпус)

Определение типов ответов

Таксономия типов ответов (из Wordnet)



Типичная архитектура QA-систем



Извлечение фрагментов текста

- IR-система возвращает список документов
 - Необходимым фрагментом может быть предложение или параграф
 - Необходимо выбрать фрагменты, потенциально содержащие ответ
1. Отсеять фрагменты не содержащие ответ
 - распознавание именованных сущностей и классификация ответов
 2. Отранжировать оставшиеся фрагменты
 - Правила, составленные вручную
 - Машинное обучение

Извлечение фрагментов текста (ранжирование)

- Признаки
 - Число именованных сущностей правильного типа в фрагменте
 - Число ключевых слов из вопроса в фрагменте
 - Наиболее длинная последовательность ключевых слов запроса в фрагменте
 - Ранг документа (IR), содержащего фрагмент
 - Плотность ключевых слов из вопроса в фрагменте
 - Пересечение N-грамм вопроса и фрагмента

Извлечение фрагментов

- Для извлечения ответа из Веба можно пропустить шаг извлечения фрагмента и использовать **сниппеты**, возвращаемые информационно-поисковыми системами

The screenshot shows a search engine interface with a yellow header bar. The search bar contains the text "что такое сниппет?". To the right of the search bar is a "Найти" button. Below the search bar, there are two checkboxes: "в найденном" and "в Москве". To the right of these checkboxes is a link "расширенный поиск". Below the search bar, there are two search results. The first result is titled "Описание сайта - Что такое сниппет?" and the second result is titled "Что такое сниппет и как его использовать".

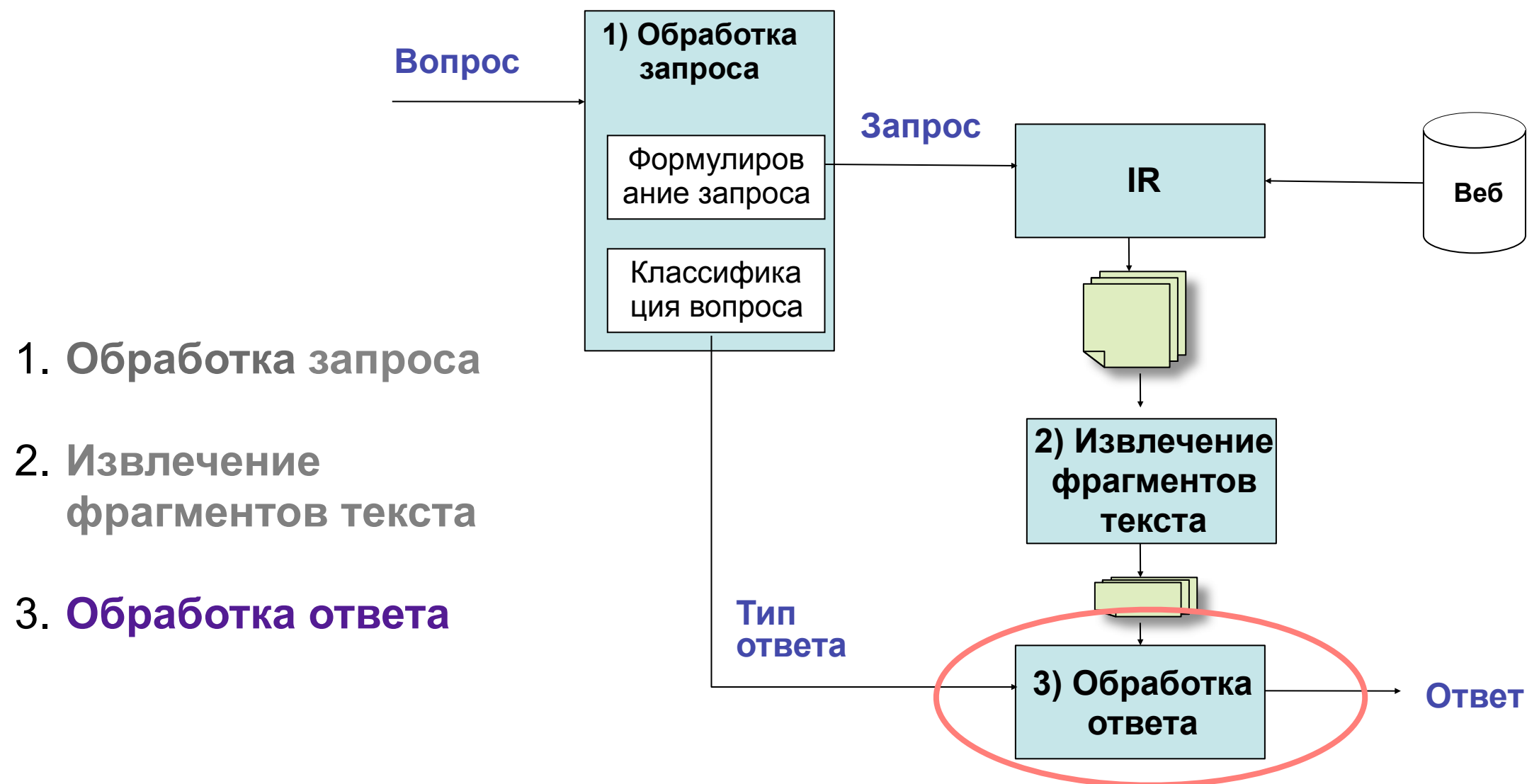
что такое сниппет? × Найти

☐ в найденном ☐ в Москве расширенный поиск

[Описание сайта - Что такое сниппет?](#)
Что представляют из себя навигационные цепочки? Для каких страниц в **сниппетах** показываются даты? Какие специальные данные могут быть показаны в **сниппетах**? **Что такое сниппет?**
[help.yandex.ru](#) > [Помощь](#) > [Вебмастер](#) [копия](#) [ещё](#)

[Что такое сниппет и как его использовать](#)
Сниппет (англ. **snippet** - лоскут, отрывок или фрагмент) - это та короткая текстовая информация по сайту, которая появляется в результатах поиска, сразу же под вылезшим адресом.
[bigfuzzy.com](#) > [Articles/Promoution...snippet.php](#) [копия](#) [ещё](#)

Типичная архитектура QA-систем



Обработка ответа

- Извлечение специфического ответа из фрагмента
- Два основных класса алгоритмов
 - Основанные на шаблонах
 - Сбор ответа из N-грамм (N-gramm tiling)

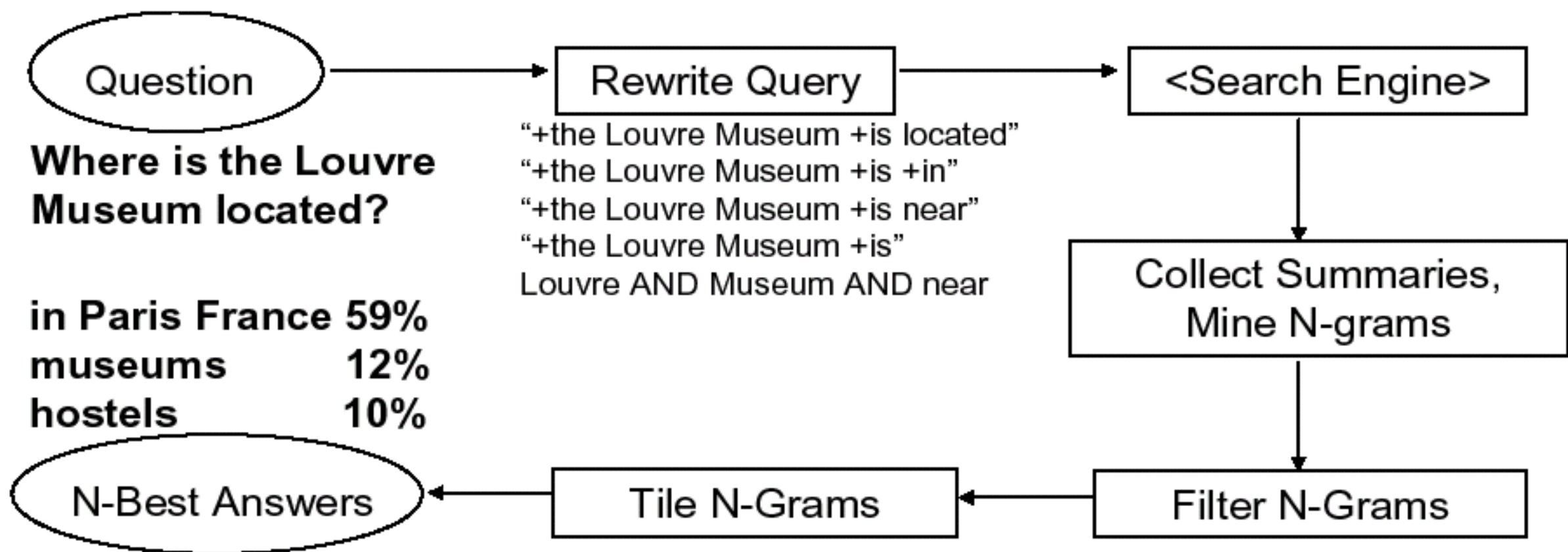
Алгоритмы на основе шаблонов

- Использование информации о типе в регулярных выражениях
 - Если тип ответа ЧЕЛОВЕК, извлечь именованные сущности ЧЕЛОВЕК из фрагмента
- Некоторые типы ответов (например, определения) не подразумевают конкретного типа именованной сущности в ответе
 - Использовать регулярные выражения (созданные вручную или автоматически)

Pattern	Question	Answer
<AP> such as <QP>	<i>What is autism?</i>	<i>developmental disorders such as autism</i>

Сбор ответа из N-грамм

Архитектура AskMSR



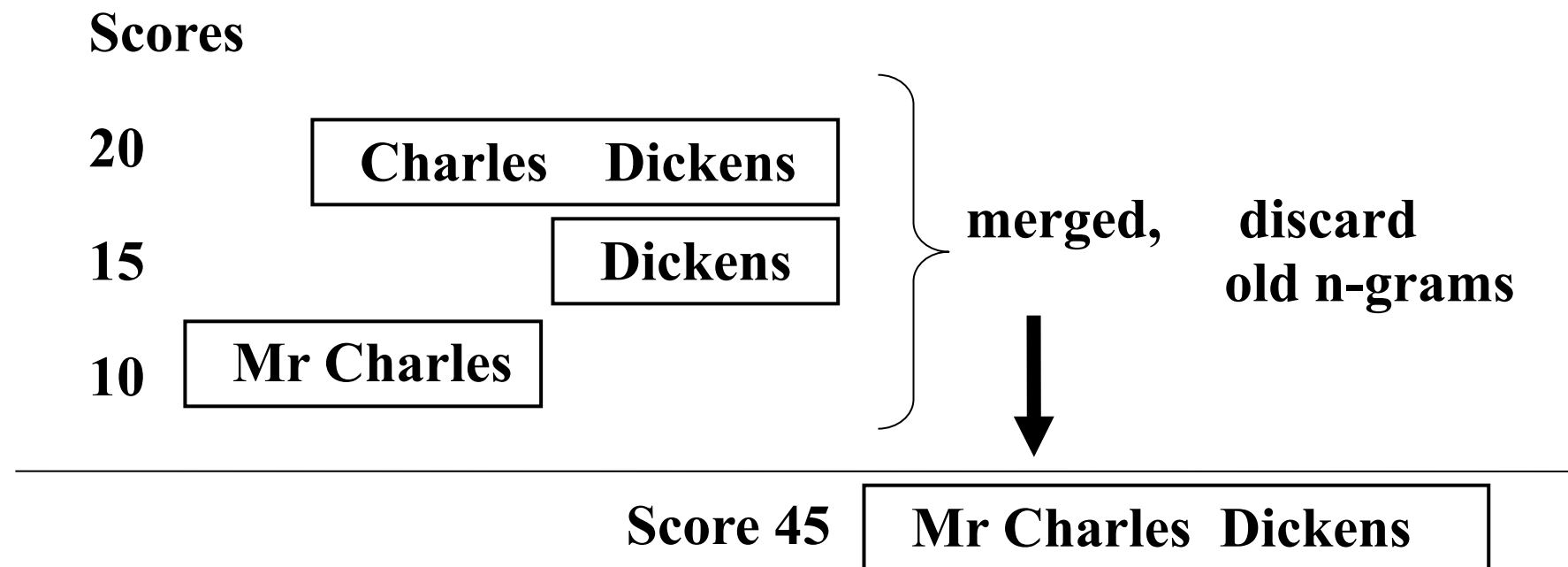
Сбор N-грамм

- Назначить вес N-грамме равный количеству снippetов, в которых она встретилась
- Пример: “Who created the character of Scrooge?”

– Dickens	117
– Christmas Carol	78
– Charles Dickens	75
– Disney	72
– Carl Banks	54
– A Christmas	41
– Christmas Carol	45
– Uncle	31

Фильтрация и сбор ответа

- Заново взвесить N-граммы с учетом типа ответа
- Собрать ответ



Автоматическое реферирование

- Часто ответом на вопрос должен быть текст
- Пример:
 - Кто такой Франсуа Томбалбай?
- Извлечение короткого фрагмента текста является задачей **автоматического реферирования**

Аннотирование и реферирование

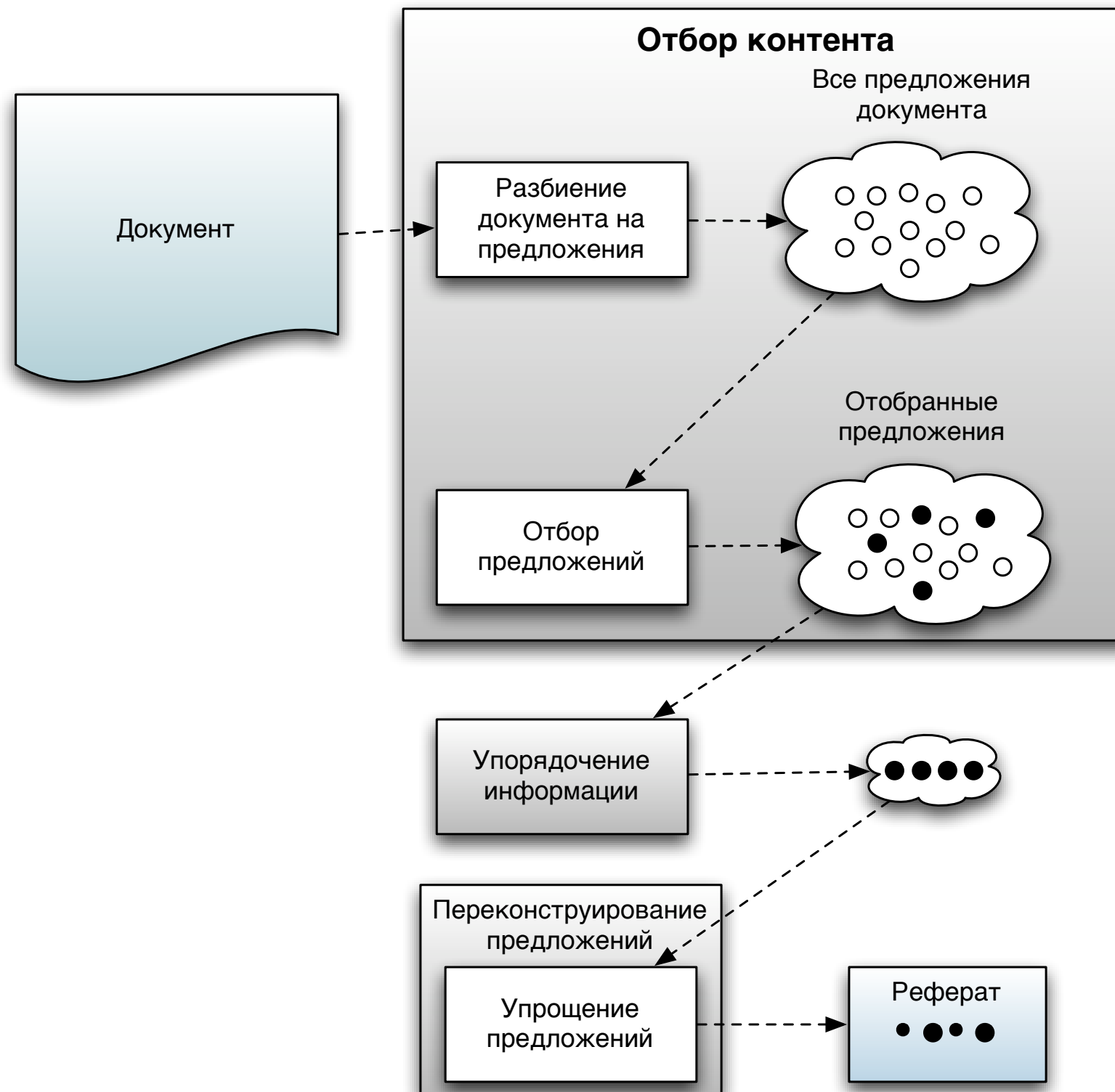
- **Реферат** состоит из частей оригинального текста
- **Аннотация** - главная мысль документа, сформулированная своими словами
 - Более компактная
 - Предполагает генерацию текста

Автоматическое реферирование

Приложения

- Аннотации и рефераты к научным и другим статьям
- Реферированное новостей (несколько документов)
- Создание сниппетов
- Текст для мобильных устройств
- Реферат встречи
- ...

Типичная архитектура



Отбор контента

- **Без учителя**

- выбор предложений с ключевыми словами (tf-idf, логарифмическое отношение правдоподобия, ...)

- Центральность

- пример $centrality(x) = \frac{1}{K} \sum_y \text{tf-idf-cos}(x, y)$

- **С учителем**

- бинарная классификация предложений

- признаки: позиция, обобщающие фразы (“in summary”, “in conclusion”, ...), информативность слов, длина предложения, связность

Упорядочение

- **Для одного документа**
 - Использовать порядок внутри документа
- **Для коллекции документов**
 - более сложные методы
 - кластеризация предложений

Переконструирование предложения

- Упрощение предложений
 - ~~When it arrives sometime new year in new TV sets, the V-chip will give parents a new and potentially revolutionary device to block out programs they don't want their children to see.~~
- Использование синтаксического разбора и удаление неинформативных частей
 - Zajic et al. 2007, Conroy et al. 2006

Следующая лекция

- Машинный перевод