ОСНОВЫ ОБРАБОТКИ ТЕКСТОВ

Лекция #9:
Другие задачи обработки текстов

Лектор: м.н.с. ИСП РАН Майоров Владимир Дмитриевич

Уже рассмотренные задачи

- Извлечение именованных сущностей
- Сегментация текста
- Морфологический анализ
- Лемматизация
- Синтаксический анализ
- Машинный перевод
- Разрешение кореферентности

Сегодня мы рассмотрим

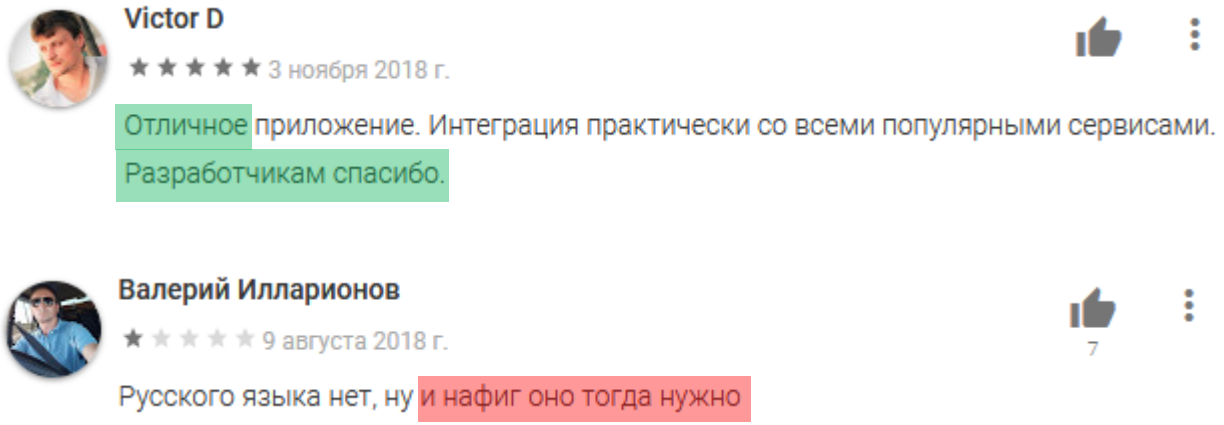
- Анализ тональности текстов
- Автоматическое реферирование текстов
- Диалоговые системы

Извлечение мнений (Opinion Mining)

- Класс задач обработки текстов, посвященных извлечению эмоционального отношения автора высказывания к объекту реального мира (предмету, событию, процессу), описанному в тексте

Анализ тональности текста

- Вход: текст (предложение)
- Выход: тональность текста
  - Положительная / отрицательная / нейтральная
  - числовая оценка тональности [-1; 1]
  - рейтинг



Анализ тональности текста

- Вход: текст (предложение)
- Выход: тональность текста
  - Положительная / отрицательная / нейтральная
  - числовая оценка тональности [-1; 1]
  - рейтинг

- Подзадачи
  - Оценка субъективности текста
  - Оценка полярности текста (для субъективных предложений)

Методы

- Словарные (словари эмоциональной лексики)
- Supervised ML
  - Признаки классификации
    - Словоформы / леммы
    - Части речи
    - Пунктуация
    - Векторные представления слов
    - …
  - Классификаторы
    - SVM
    - Naïve Bayes
    - MLP
    - Деревья решений
    - …

Словари оценочных слов

- Общие оценочные слова
  - Хороший / Плохой / Замечательный …
- Домено-специфичные оценочные слова
  - Большой / маленький
  - Тихий / громкий
  - …
- Слова модификаторы
  - Слишком / чересчур / недостаточно …

Извлечение мнений

- Вход: текст
- Выход: множество мнений

Мнение – $< Holder, Target, Sentiment, Time >$

- $Holder$ – тот, кто мнение выражает
- $Target$ – цель мнения
  - $Object$ - объект
  - $Feature\ (aspect)$ – характеристика объекта (или *overall*)
- $Sentiment$ – тональность мнения
- $Time$ – время выражения мнения

Извлечение мнений

Мнение – $< Holder, Target, Sentiment, Time >$

- Расположение
- Обслуживание
- Интерьер
- Кухня
- Бар

Elena B
249 94

Отзыв написан 20 октября 2017 г.

Очень даже!

Ресторан понравился.сидели на летней площадке.отличное обслуживание.прекрасно оформленные блюда.вкусно и порции достаточно немаленькие.вкусные десерты.хорошая винная карта.из всех ресторанов сан-Себастьяна -этот занял первое место!бронировать обязательно.ехать за город,но везде стоят указатели,так что найти не составило труда

Спасибо, Elena B!

Извлечение мнений

Мнение – $< Holder, Target, Sentiment, Time >$

- Расположение
- Обслуживание
- Интерьер
- Кухня
- Бар

Elena B
249 94

Отзыв написан 20 октября 2017 г.

Очень даже!

Ресторан понравился.сидели на летней площадке.отличное обслуживание.прекрасно оформленные блюда.вкусно и порции достаточно немаленькие.вкусные десерты.хорошая винная карта.из всех ресторанов сан-Себастьяна -этот занял первое место!бронировать обязательно.ехать за город,но везде стоят указатели,так что найти не составило труда

Спасибо, Elena B!

<Elena B, (Ресторан, overall), +, 20.10.2017>

Извлечение мнений

Мнение – $< Holder, Target, Sentiment, Time >$

- Расположение
- Обслуживание
- Интерьер
- Кухня
- Бар

Elena B
249 94

Отзыв написан 20 октября 2017 г.

Очень даже!

Ресторан понравился.сидели на летней площадке.отличное обслуживание.прекрасно оформленные блюда.вкусно и порции достаточно немаленькие.вкусные десерты.хорошая винная карта.из всех ресторанов сан-Себастьяна -этот занял первое место!бронировать обязательно.ехать за город,но везде стоят указатели,так что найти не составило труда

Спасибо, Elena B!

<Elena B, (Ресторан, overall), +, 20.10.2017>

<Elena B, (Ресторан, обслуживание), +, 20.10.2017>

Извлечение мнений

Мнение – $< Holder, Target, Sentiment, Time >$

- Расположение
- Обслуживание
- Интерьер
- Кухня
- Бар

Elena B
249 94

Отзыв написан 20 октября 2017 г.

Очень даже!

Ресторан понравился.сидели на летней площадке.отличное обслуживание.прекрасно оформленные блюда.вкусно и порции достаточно немаленькие.вкусные десерты.хорошая винная карта.из всех ресторанов сан-Себастьяна -этот занял первое место!бронировать обязательно.ехать за город,но везде стоят указатели,так что найти не составило труда

Спасибо, Elena B!

<Elena B, (Ресторан, overall), +, 20.10.2017>

<Elena B, (Ресторан, обслуживание), +, 20.10.2017>

<Elena B, (Ресторан, кухня), +, 20.10.2017>

Извлечение мнений

Мнение – $< Holder, Target, Sentiment, Time >$

- Расположение
- Обслуживание
- Интерьер
- Кухня
- Бар

Elena B
249 94

Отзыв написан 20 октября 2017 г.

Очень даже!

Ресторан понравился.сидели на летней площадке.отличное обслуживание.прекрасно оформленные блюда.вкусно и порции достаточно немаленькие.вкусные десерты.хорошая винная карта.из всех ресторанов сан-Себастьяна -этот занял первое место!бронировать обязательно.ехать за город,но везде стоят указатели,так что найти не составило труда

Спасибо, Elena B!

<Elena B, (Ресторан, overall), +, 20.10.2017>

<Elena B, (Ресторан, обслуживание), +, 20.10.2017>

<Elena B, (Ресторан, кухня), +, 20.10.2017>

<Elena B, (Ресторан, бар), +, 20.10.2017>

Извлечение мнений

Мнение – $< Holder, Target, Sentiment, Time >$

- Расположение
- Обслуживание
- Интерьер
- Кухня
- Бар

Elena B
249 94

Отзыв написан 20 октября 2017 г.

Очень даже!

Ресторан понравился.сидели на летней площадке.отличное обслуживание.прекрасно оформленные блюда.вкусно и порции достаточно немаленькие.вкусные десерты.хорошая винная карта.из всех ресторанов сан-Себастьяна -этот занял первое место!бронировать обязательно.ехать за город,но везде стоят указатели,так что найти не составило труда

Спасибо, Elena B!

<Elena B, (Ресторан, overall), +, 20.10.2017>

<Elena B, (Ресторан, обслуживание), +, 20.10.2017>

<Elena B, (Ресторан, кухня), +, 20.10.2017>

<Elena B, (Ресторан, бар), +, 20.10.2017>

<Elena B, (Ресторан, расположение), +, 20.10.2017>

Подзадачи

- Извлечение объектов мнения
  задача похожа на mention detection
- Извлечение характеристик объектов
  - Извлечение явных упоминаний характеристик
    задача похожа на NERC
  - Извлечение неявных упоминаний характеристик
    multiclass классификация предложений/клауз
- Оценка тональности по отношению к объекту/характеристике объекта

Оценка тональности по отношению к объекту

- Также как и в анализе тональности предложений применяются методы supervised ML

- Во многих методах признаки основаны на синтаксическом дереве предложений
  - Слова на пути от оценочных слов до упоминания объекта
  - Слова на пути от корня до упоминания объекта
  - Зависимые слова упоминания объекта
  - …

Анализ тональности текста

- Основные проблемы методов
  - Обработка отрицаний
    - «Мне не очень нравится …», …
  - Сравнение объектов
    - «из всех ресторанов сан-Себастьяна -этот занял первое место!»
  - Противоречивая оценка
    - «С одной стороны …, с другой …»
  - Ирония/сарказм

Анализ тональности текста

- Методы оценки качества

$$Precision = \frac{correct}{predicted}$$

$$Recall = \frac{correct}{expected}$$

$$F_1 = \frac{2\ Precision\ Recall}{Precision + Recall}$$

Реферирование

- Реферирование – процесс извлечения из текста основного содержания и заданной информации с целью их письменного изложения (в виде реферата)

- Реферат – краткое точное изложение содержания исходного документа.

Реферирование

- Реферирование – процесс извлечения из текста основного содержания и заданной информации с целью их письменного изложения (в виде реферата)

- Реферат – краткое точное изложение содержания исходного документа.

- Задача: по тексту построить более короткий текст (реферат), передающий основное содержание исходного текста

Реферирование

- Варианты задачи:
  - Extraction-based

    Реферат состоит из неизмененных (или незначительно измененных) фрагментов исходного текста
  - Abstraction-based:

    Предложения (слова) реферата не содержатся в исходном тексте.

  - Single-document

    реферат строится по 1 тексту
  - Multi-document:

    реферат строится по коллекции близких по смыслу текстов

Реферирование

- Методы решения задачи
  - Статистические методы
  - Графовые методы
  - Дискурсные методы

Статистические Методы

- Общий подход к решению
  - Сегментация текста;
    (предложения/клаузы/словосочетания);
  - Предобработка;
    (удаление пунктуации, стоп-слов, нормализация слов)
  - Оценка важности предложения по признакам;
  - Выбор наиболее важных сегментов;
  - Упорядочивание сегментов и формирование реферата.

Статистические Методы

- Признаки
  - На уровне слова
    - Именованные сущности
    - Числа, даты
    - TF*IDF
    - …
  - На уровне предложения
    - Длина (относительная)
    - Расстояние от начала документа/параграфа
    - Похожесть на заголовок текста
    - Похожесть на соседние предложения
    - …

Графовые методы

- Общий подход к решению
  - Сегментация текста;
    (предложения/клаузы/словосочетания);
  - Построение графа исходных текстов
    - Узлы графа – выделенные сегменты
    - Дуги графа – отношения похожести между сегментами
  - Кластеризация графа
  - Выбор наиболее представительных узлов кластеров
  - Упорядочивание сегментов и формирование реферата.

Оценка качества

- Ручные методы
  - Связность
  - Грамматическая правильность
  - Избыточность
  - Полнота
- Полуавтоматические методы
  - Pyramid
- Автоматические методы
  - Precision, Recall, F1
  - ROUGE

ROUGE-N (N-gram Co-Occurrence Statistics)

$$ROUGE - N = \frac{\sum_{S \in Gold} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in Gold} \sum_{gram_n \in S} Count(gram_n)}$$

ROUGE-L (Longest Common Subsequence)

- Последовательность $Z = [z_1, z_2, \dots, z_n]$ является подпоследовательностью $X = [x_1, x_2, \dots, x_m]$, если $\exists [i_1, i_2, \dots i_n]\ (i_k < i_l, k < l)$: $x_{i_j} = z_j, j = \overline{1,n}$

- $LCS(X, Y)$ – длина наибольшей общей подпоследовательности

- Метрики на уровне предложений:
  - $P_{lcs} = \frac{LCS(G,S)}{|S|}, R_{lcs} = \frac{LCS(G,S)}{|G|}, F_{lcs} = \frac{2R_{lcs}P_{lcs}}{R_{lcs}+P_{lcs}};$
- Метрики на уровне реферата:
  - $P_{lcs} = \frac{\sum_i LCS_\cup(g_i,S)}{|S|}, R_{lcs} = \frac{\sum_i LCS_\cup(g_i,S)}{|G|}, F_{lcs} = \frac{2R_{lcs}P_{lcs}}{R_{lcs}+P_{lcs}};$

Другие метрики ROUGE

- ROUGE-W (Weighted Longest Common Sub-sequence)
  - Похожа на ROUGE-L, но дополнительно учитывает длины последовательных совпадений
- ROUGE-S (Skip-Bigram Co-Occurrence Statistics)
  - Похожа на ROUGE-N, но оценивает количество общих skip-bigram
- ROUGE-SU (Extension of ROUGE-S)
  - ROUGE-S с учетом пересечения униграмм

Варианты постановок задач

- Extraction-based vs Abstraction-based
- Single-document vs Multi-document

- Generic vs Query-focused
- Mono-lingual vs Multi-lingual vs Cross-lingual

- Update summarization

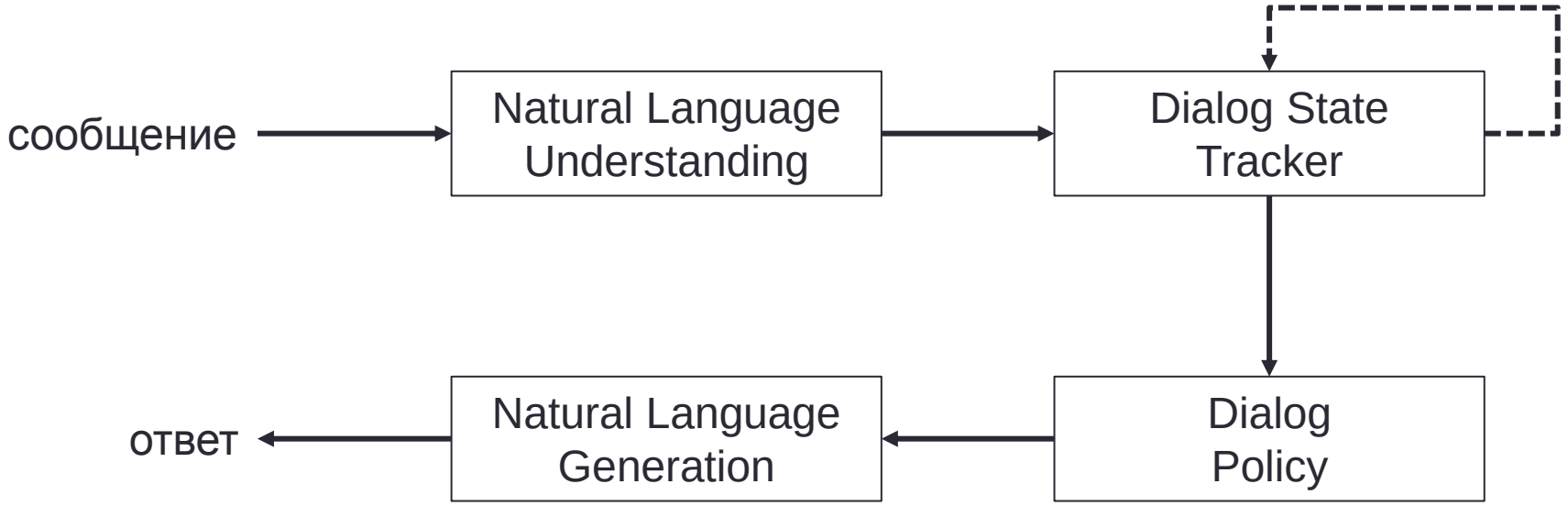
Диалоговые системы

- Диалог – это процесс обмена сообщениями между пользователем и компьютером.

- Виды диалогов
  - Вопрос-ответ
  - Диалоги с определенной целью (решение конкретной задачи)
  - Диалоги без цели

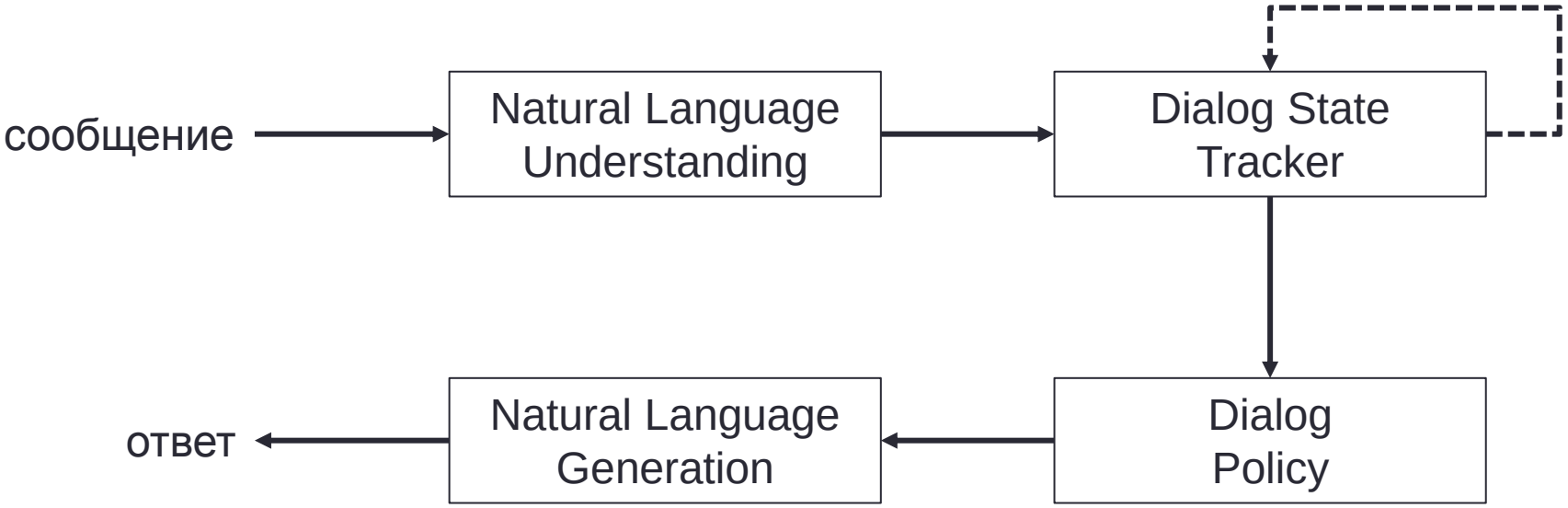
Диалоговые системы (с целью)

- Общая схема решения:



Диалоговые системы (с целью)

- Общая схема решения:



– Хочу забронировать билет на самолет в Париж в 17:00

Диалоговые системы (с целью)

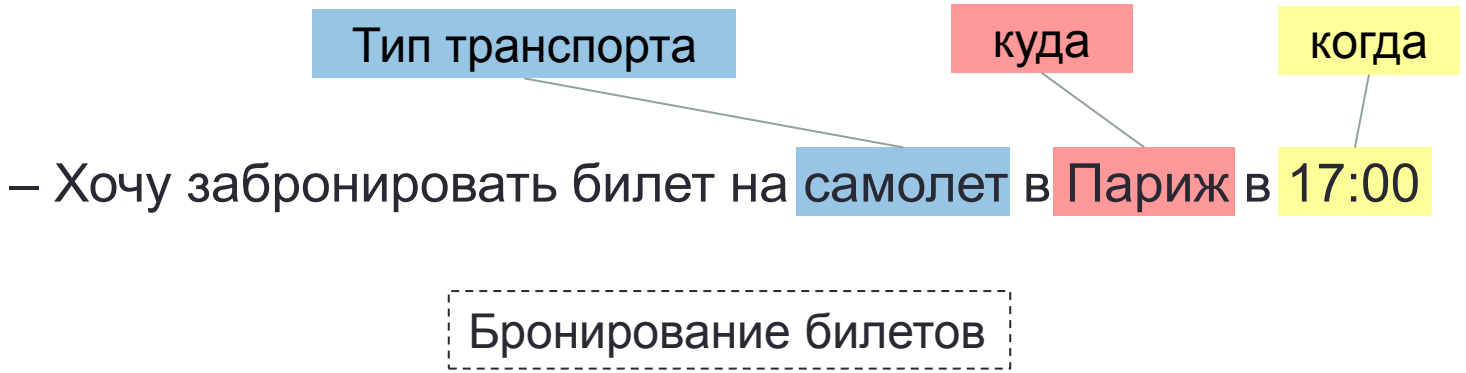
- Natural Language Understanding
  - Определяется намерение пользователя
    - Решается как задача классификации текста

– Хочу забронировать билет на самолет в Париж в 17:00

Бронирование билетов

Диалоговые системы (с целью)

- Natural Language Understanding
  - Определяется намерение пользователя
    - Решается как задача классификации текста
  - Определяются значения полей
    - NERC
    - Решается как задача разметки последовательности слов



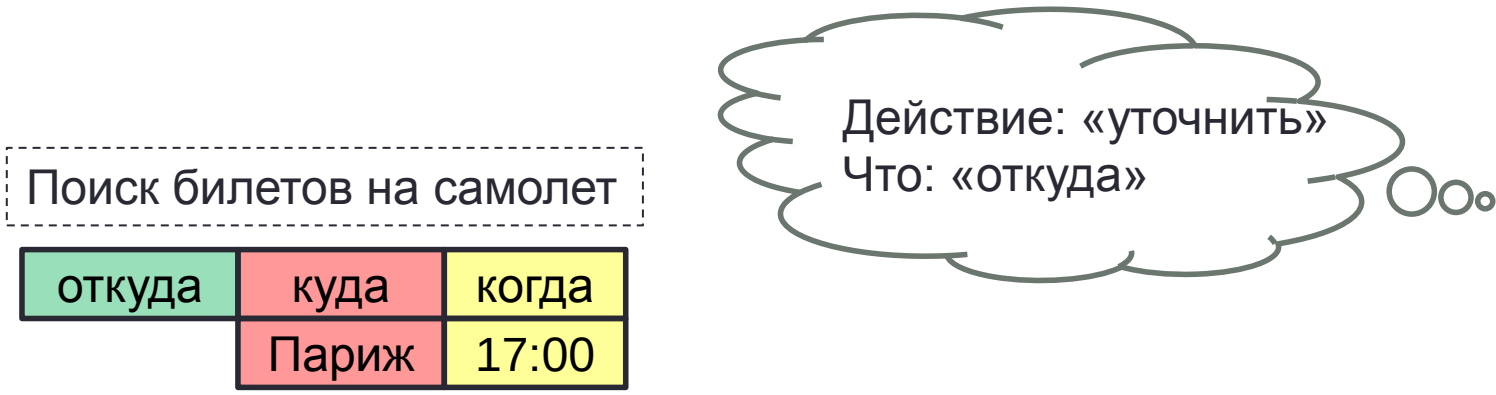
Диалоговые системы (с целью)

- Dialog State Tracker
  - Определяет цель пользователя (с учетом прошлых состояний)
    - Правила / supervised ML
  - Формирует запрос (заполняет «форму» полями)
    - Правила / supervised ML

– Хочу забронировать билет на самолет в Париж в 17:00

Диалоговые системы (с целью)

- Dialog Policy
  - На основе текущего состояния и «формы» выполняет некоторое действие



– Хочу забронировать билет на самолет в Париж в 17:00

Диалоговые системы (с целью)

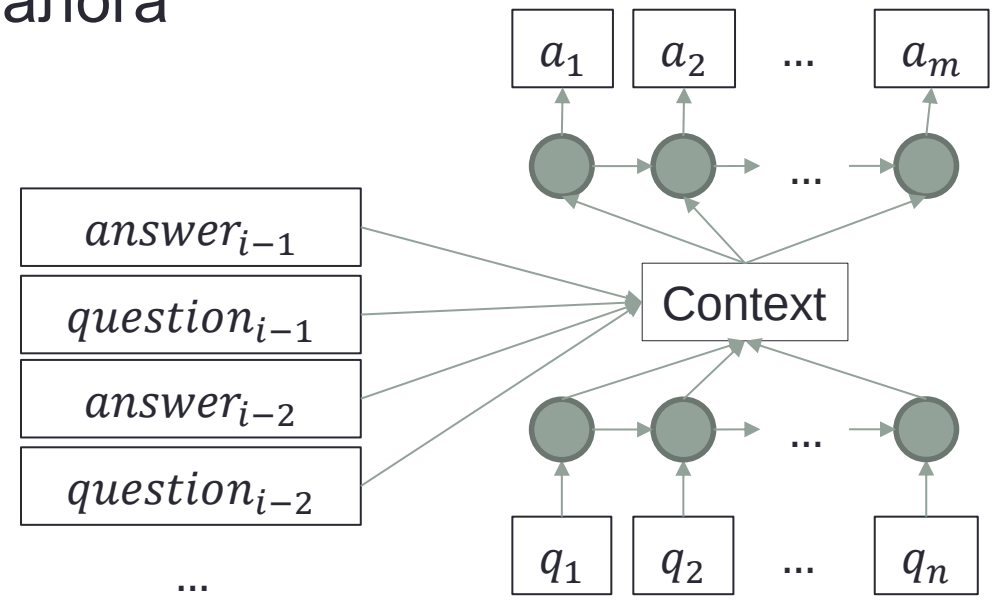
- Natural Language Generation
  - Формирует представление результата действия в виде текста на естественном языке
    - Шаблоны ответов
    - Готовые ответы из списка

– Хочу забронировать билет на самолет в Париж в 17:00
– Уточните, откуда Вы собираетесь вылетать.

Диалоговые системы (без цели)

- В основном решаются как задача машинного перевода
  - Сообщение пользователя – текст на исходном «языке»
  - Ответ системы – текст на целевом языке

- В модели кодировщик-декодировщик также может учитываться история диалога



Следующая лекция

- Извлечение отношений

- Лектор: Андрианов Иван Алексеевич