



NLU: основные понятия и инструменты

Сердюк Даша,
инженер-разработчик Just AI, отдел
разработки решений, группа разработки
голосовых ассистентов

Как мы проведем сегодня вечер

- обсудим NLU-модуль диалоговых систем, их основные понятия и инструменты

Как мы проведем сегодня вечер

- обсудим NLU-модуль диалоговых систем, их основные понятия и инструменты
- лекция на ~1,5 часа по теме “Структура диалоговых систем (NLU/DM/NLG). NLP задачи в разработке диалоговых систем. Подходы и алгоритмы в распознавании интенгов и именованных сущностей”



Как мы проведем сегодня вечер

- обсудим NLU-модуль диалоговых систем, их основные понятия и инструменты
- лекция на ~1,5 часа по теме “Структура диалоговых систем (NLU/DM/NLG). NLP задачи в разработке диалоговых систем. Подходы и алгоритмы в распознавании интенгов и именованных сущностей”
- перерыв на 10 минут

Как мы проведем сегодня вечер

- обсудим NLU-модуль диалоговых систем, их основные понятия и инструменты
- лекция на ~1,5 часа по теме “Структура диалоговых систем (NLU/DM/NLG). NLP задачи в разработке диалоговых систем. Подходы и алгоритмы в распознавании интенгов и именованных сущностей”
- перерыв на 10 минут
- лекция на ~1,5 часа по теме “Обзор инструментов разработки диалоговых систем: визуальные конструкторы, платформы и фреймворки”

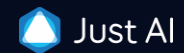
Как мы проведем сегодня вечер

- обсудим NLU-модуль диалоговых систем, их основные понятия и инструменты
- лекция на ~1,5 часа по теме “Структура диалоговых систем (NLU/DM/NLG). NLP задачи в разработке диалоговых систем. Подходы и алгоритмы в распознавании интенгов и именованных сущностей”
- перерыв на 10 минут
- лекция на ~1,5 часа по теме “Обзор инструментов разработки диалоговых систем: визуальные конструкторы, платформы и фреймворки”
- вопросы - пишите в чат, после смысловых блоков по ходу рассказа буду отвечать

Как мы проведем сегодня вечер

- обсудим NLU-модуль диалоговых систем, их основные понятия и инструменты
- лекция на ~1,5 часа по теме “Структура диалоговых систем (NLU/DM/NLG). NLP задачи в разработке диалоговых систем. Подходы и алгоритмы в распознавании интенгов и именованных сущностей”
- перерыв на 10 минут
- лекция на ~1,5 часа по теме “Обзор инструментов разработки диалоговых систем: визуальные конструкторы, платформы и фреймворки”
- вопросы - пишите в чат, после смысловых блоков по ходу рассказа буду отвечать
- будет домашнее задание, в проверяющих ставьте Дарью Сердюк и Светлану Вольскую

Обо мне



just-ai.com



Кто такая эта Даша Сердюк?

- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки



Кто такая эта Даша Сердюк?

- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки
- 6+ лет в NLP

Кто такая эта Даша Сердюк?

- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки
- 6+ лет в NLP
- 5+ лет в Conversational AI

Кто такая эта Даша Сердюк?

- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки
- 6+ лет в NLP
- 5+ лет в Conversational AI
- Премия Алисы

Кто такая эта Даша Сердюк?

- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки
- 6+ лет в NLP
- 5+ лет в Conversational AI
- Премия Алисы
- автор курсов и обучающих материалов по CAI

Кто такая эта Даша Сердюк?

- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки
- 6+ лет в NLP
- 5+ лет в Conversational AI
- Премия Алисы
- автор курсов и обучающих материалов по CAI
- разработчица чат-ботов

Кто такая эта Даша Сердюк?

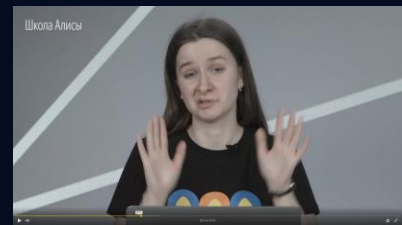
- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки
- 6+ лет в NLP
- 5+ лет в Conversational AI
- Премия Алисы
- автор курсов и обучающих материалов по CAI
- разработчица чат-ботов
- NLP Research Engineer

Кто такая эта Даша Сердюк?

- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки
- 6+ лет в NLP
- 5+ лет в Conversational AI
- Премия Алисы
- автор курсов и обучающих материалов по CAI
- разработчица чат-ботов
- NLP Research Engineer
- снова разработчица, но уже комплексных решений для голосовых ассистентов :)

Кто такая эта Даша Сердюк?

- CS в бакалавриате в ПГНИУ, DS в магистратуре Питерской Вышки
- 6+ лет в NLP
- 5+ лет в Conversational AI
- Премия Алисы
- автор курсов и обучающих материалов по CAI
- разработчица чат-ботов
- NLP Research Engineer
- снова разработчица, но уже комплексных решений для голосовых ассистентов :)



Начинаем!





Взгляд издалека

По цели использования и по предметной области (домену)

Взгляд издалека

По цели использования и по предметной области (домену)

- целеориентированные



Взгляд издалека

По цели использования и по предметной области (домену)

- целеориентированные
- ДС общей направленности (цель на "поболтать")

Взгляд издалека

По цели использования и по предметной области (домену)

- целеориентированные
- ДС общей направленности (цель на "поболтать")

Разделение ДС по предметной области

Взгляд издалека

По цели использования и по предметной области (домену)

- целеориентированные
- ДС общей направленности (цель на "поболтать")

Разделение ДС по предметной области

- закрытого домена (в первом случае чатбот "знает" только о какой-то определенной теме/домене)



Взгляд издалека

По цели использования и по предметной области (домену)

- целеориентированные
- ДС общей направленности (цель на "поболтать")

Разделение ДС по предметной области

- закрытого домена
- открытого домена (ожидаем от чатбота, что он сможет поддерживать разговор на <почти> любую тему)

Взгляд издалека

По цели использования и по предметной области (домену)

- целеориентированные
- ДС общей направленности

Голосовые ассистенты могут совмещать в себе несколько чатботов разных по предметной области и цели использования!

Разделение ДС по предметной области

- закрытого домена
- открытого домена

время посмотреть вопросы в чатике...

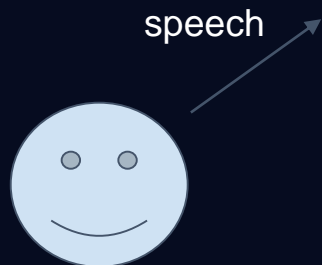
Строение диалоговых систем



NLU

(Natural
Language
Understanding)

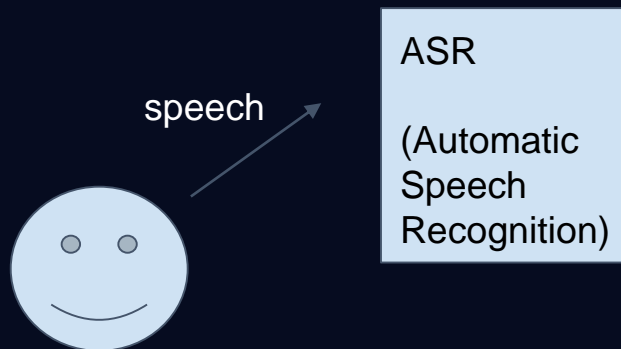
Строение диалоговых систем



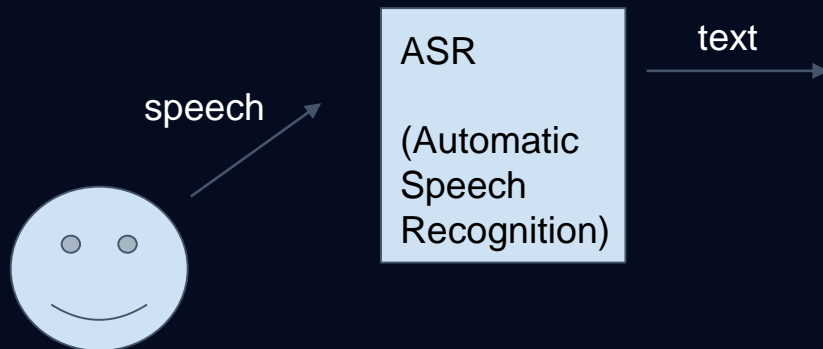
NLU

(Natural
Language
Understanding)

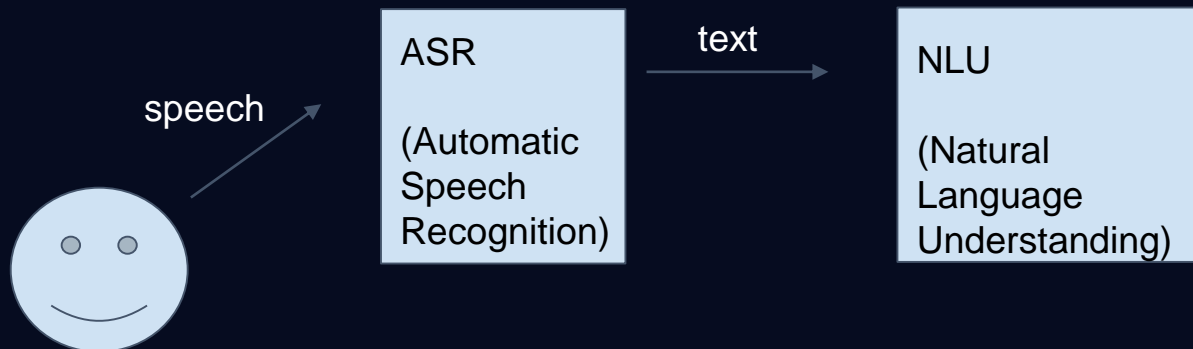
Строение диалоговых систем



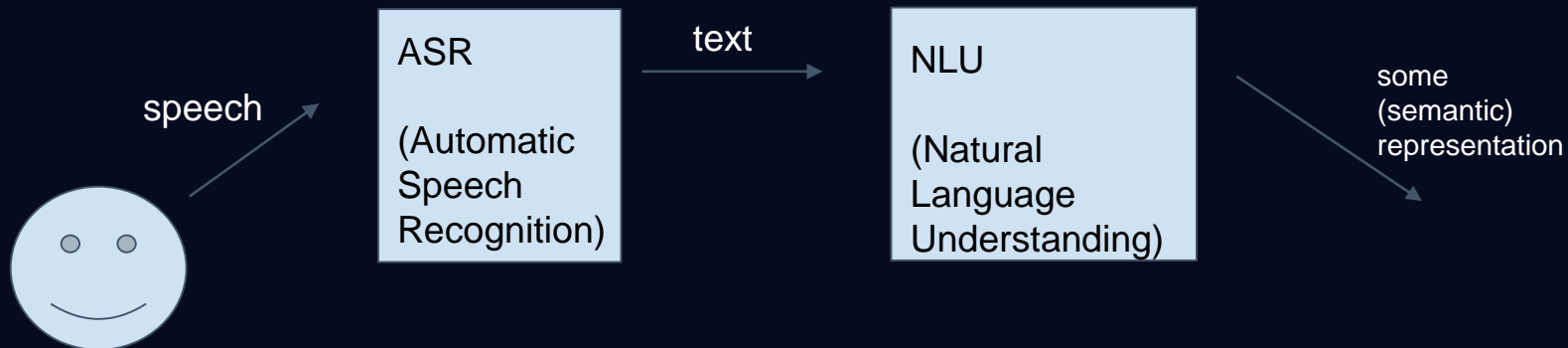
Строение диалоговых систем



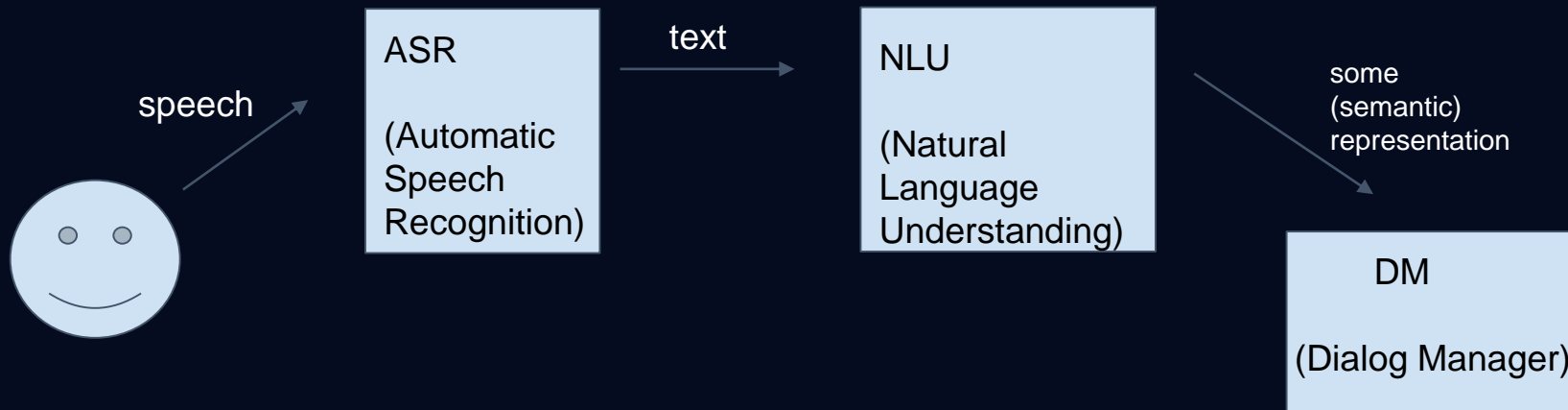
Строение диалоговых систем



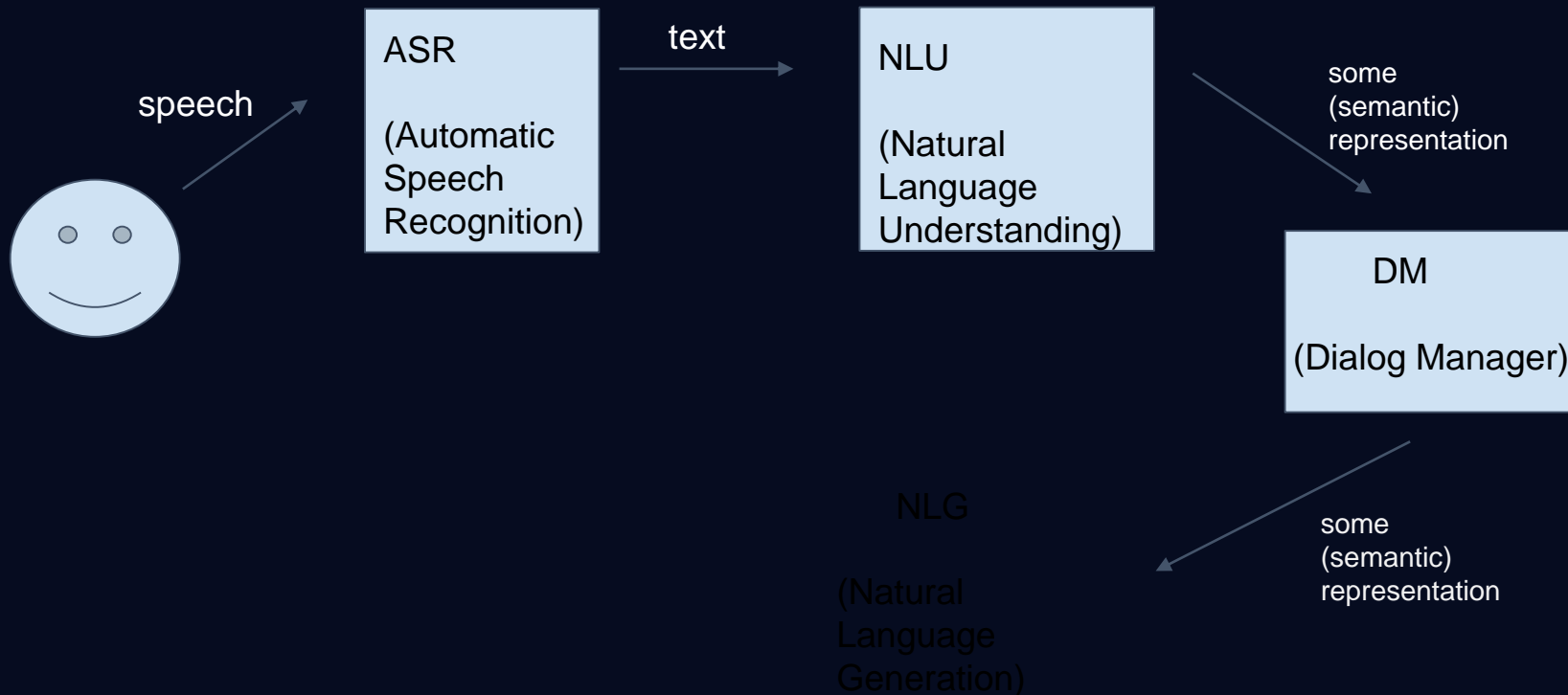
Строение диалоговых систем



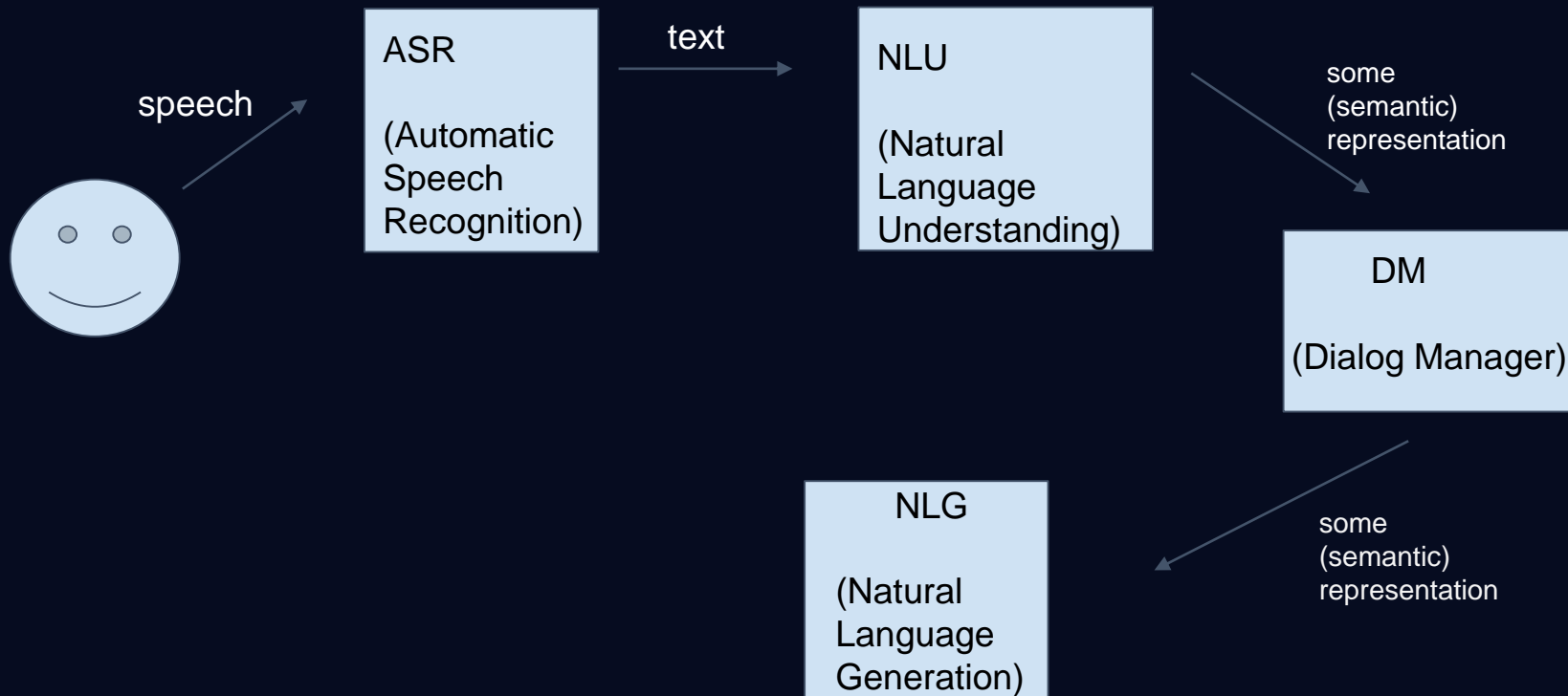
Строение диалоговых систем



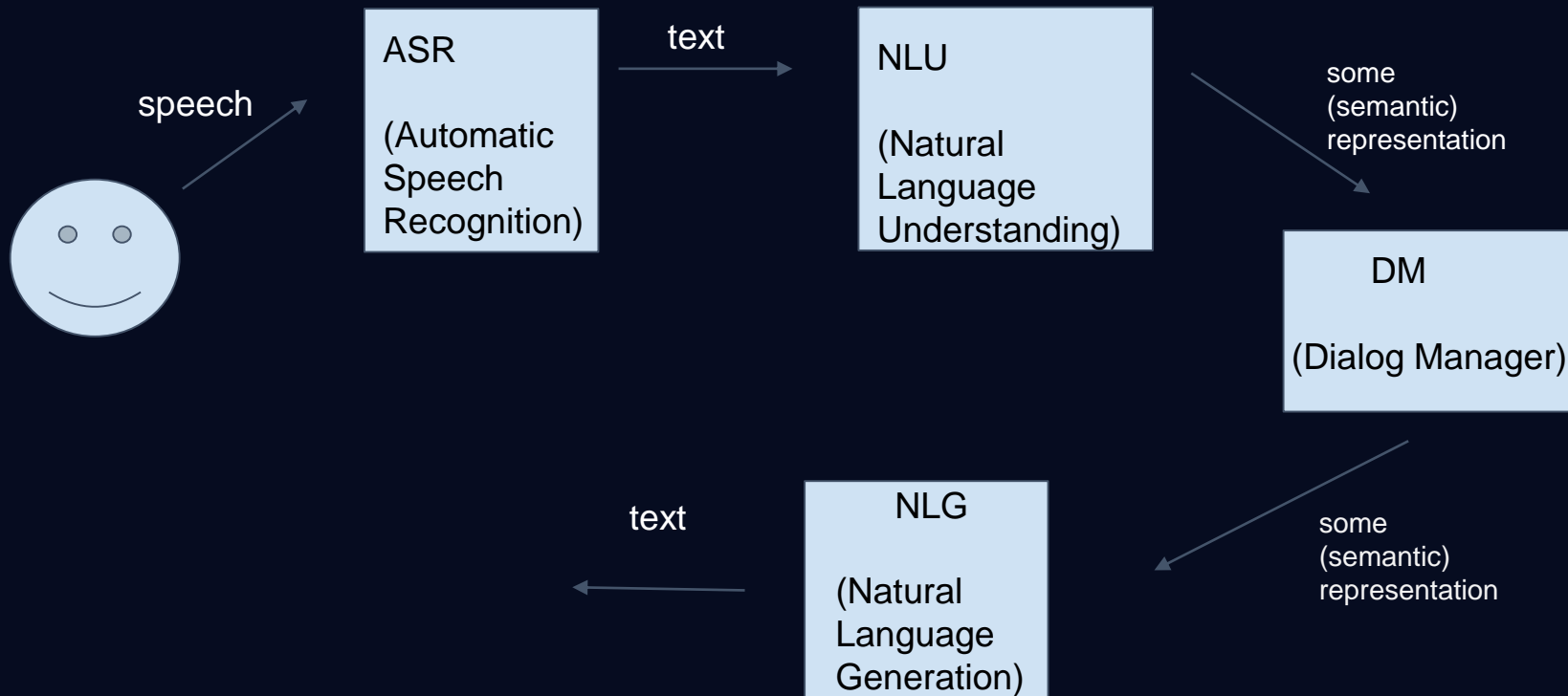
Строение диалоговых систем



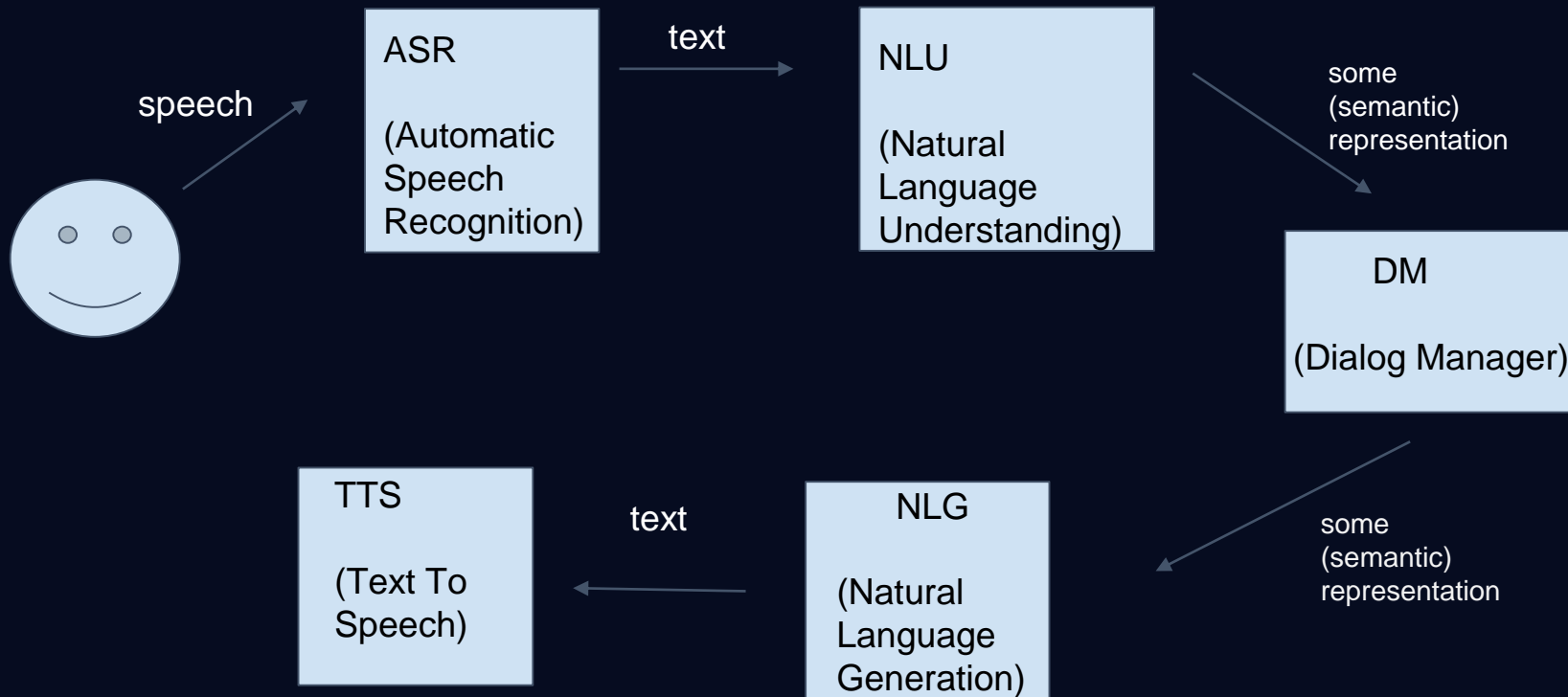
Строение диалоговых систем



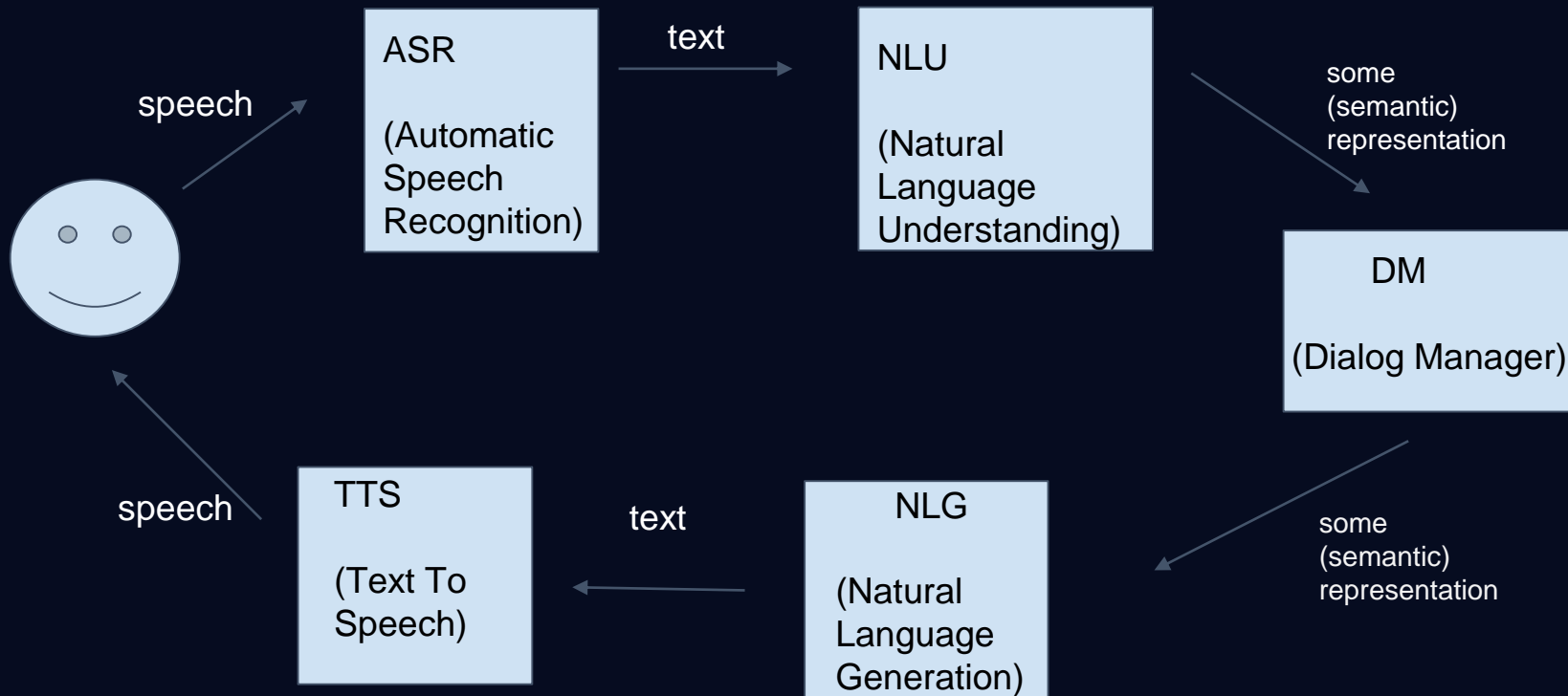
Строение диалоговых систем



Строение диалоговых систем



Строение диалоговых систем



Dialog Manager

$U \rightarrow ASR \rightarrow NLU \rightarrow (\quad DM \quad) \rightarrow NLG \rightarrow TTS \rightarrow U$

$U \rightarrow ASR \rightarrow NLU \rightarrow (DST + DPL) \rightarrow NLG \rightarrow TTS \rightarrow U$

DST (Dialog State Tracker)

DPL (Dialog Policy Learning)

Dialog Manager

$U \rightarrow ASR \rightarrow NLU \rightarrow (DST + DPL) \rightarrow NLG \rightarrow TTS \rightarrow U$

DST (Dialog State Tracker)

определяет то, что пользователь хочет в данный момент диалога:
цель, слоты и др.



Dialog Manager

$U \rightarrow ASR \rightarrow NLU \rightarrow (DST + DPL) \rightarrow NLG \rightarrow TTS \rightarrow U$

DPL (Dialog Policy Learning)

определяет каким образом диалоговая система ответит
пользователю

(важно: для целеориентированных!)

общая схема: $U \rightarrow ASR \rightarrow NLU \rightarrow DM \rightarrow NLG \rightarrow TTS \rightarrow U$

общая схема: $U \rightarrow ASR \rightarrow NLU \rightarrow DM \rightarrow NLG \rightarrow TTS \rightarrow U$

если бот текстовый: $U \rightarrow NLU \rightarrow DM \rightarrow NLG \rightarrow U$

общая схема: $U \rightarrow ASR \rightarrow NLU \rightarrow DM \rightarrow NLG \rightarrow TTS \rightarrow U$

если бот текстовый: $U \rightarrow NLU \rightarrow DM \rightarrow NLG \rightarrow U$

если текстовый бот end-to-end: $U \rightarrow \text{end2end} \rightarrow U$

общая схема: $U \rightarrow ASR \rightarrow NLU \rightarrow DM \rightarrow NLG \rightarrow TTS \rightarrow U$

если бот текстовый: $U \rightarrow NLU \rightarrow DM \rightarrow NLG \rightarrow U$

если текстовый бот end-to-end: $U \rightarrow \text{end2end} \rightarrow U$

если бот end-to-end: $U \rightarrow ASR \rightarrow \text{end2end} \rightarrow TTS \rightarrow U$

Зачем все эти разделения и связи между модулями?

Зная, какой структуры будет диалоговая система, можно наиболее оптимально и удачно подобрать инструменты и алгоритмы для разработки.

время посмотреть вопросы в чатике...

Цель NLU-модуля

ВЫЖАТЬ все самое полезное из фразы пользователя



Основные задачи NLP

Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)

Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)
- **на уровне слов:** error correction (исправление опечаток), normalization (нормализация: лемматизация или стемминг), morphological analysis (морфологический анализ)

“собаки” → “собака”
“думали” → “думать”

“яблочный” → “яблочн”
“думали” → “дума”

ба^льшой к^рсивый оф^ес -> бо^льшой к^расивый оф^ис

```
>>> p = morph.parse('стали')[0]
>>> p.tag
OpencorporaTag('VERB,perf,intr plur,past,indc')
```

Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)
- **на уровне слов:** error correction (исправление опечаток), normalization (нормализация: лемматизация или стемминг), morphological analysis (морфологический анализ)
- **на уровне словосочетаний:** POS-tagging (разметка частей речи), named entity recognition (извлечение именованных сущностей), word segmentation (сегментация слов)

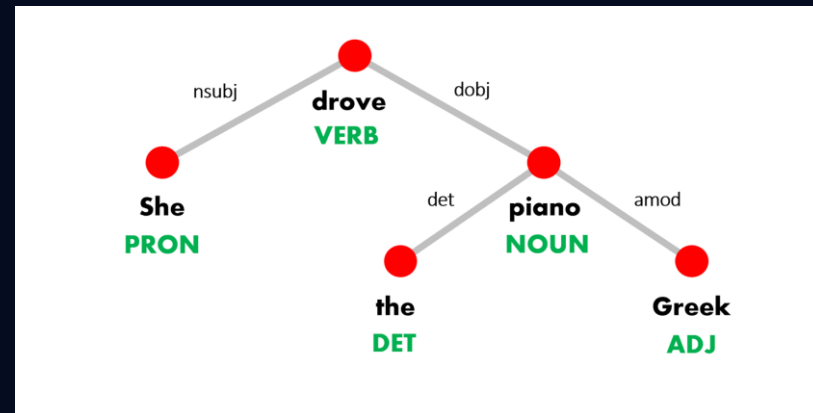
words = [James, Blake, приезжал, в, Россию, летом, 2019, года]

James Blake приезжал в Россию летом 2019 года

- musician: James Blake
- location: Россия
- date: лето 2019 года

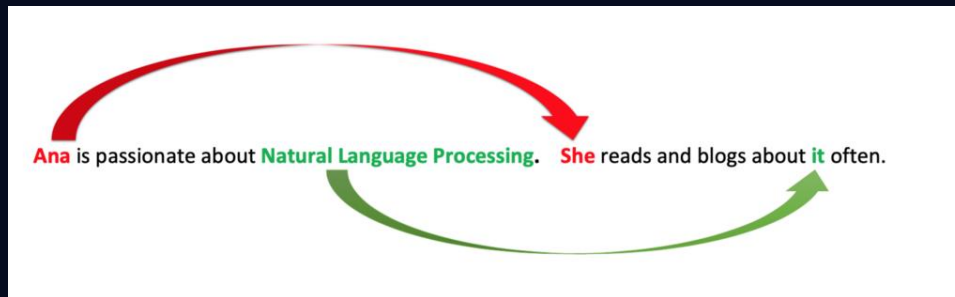
Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)
- **на уровне слов:** error correction (исправление опечаток), normalization (нормализация: лемматизация или стемминг), morphological analysis (морфологический анализ)
- **на уровне словосочетаний:** POS-tagging (разметка частей речи), named entity recognition (извлечение именованных сущностей), word segmentation (сегментация слов)
- **на уровне предложений:** sentence segmentation (сегментация предложений), query answering (ответы на вопросы), word sense disambiguation (снятие лексической многозначности), syntax parsing (синтаксический разбор)



Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)
- **на уровне слов:** error correction (исправление опечаток), normalization (нормализация: лемматизация или стемминг), morphological analysis (морфологический анализ)
- **на уровне словосочетаний:** POS-tagging (разметка частей речи), named entity recognition (извлечение именованных сущностей), word segmentation (сегментация слов)
- **на уровне предложений:** sentence segmentation (сегментация предложений), query answering (ответы на вопросы), word sense disambiguation (снятие лексической многозначности), syntax parsing (синтаксический разбор)
- **на уровне параграфов:** coreference resolution (разрешение кореференций), language detection (определение языка), sentiment analysis (анализ тональности), intent recognition (распознавание интенгов)



Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)
- **на уровне слов:** error correction (исправление опечаток), normalization (нормализация: лемматизация или стемминг), morphological analysis (морфологический анализ)
- **на уровне словосочетаний:** POS-tagging (разметка частей речи), named entity recognition (извлечение именованных сущностей), word segmentation (сегментация слов)
- **на уровне предложений:** sentence segmentation (сегментация предложений), query answering (ответы на вопросы), word sense disambiguation (снятие лексической многозначности), syntax parsing (синтаксический разбор)
- **на уровне параграфов:** coreference resolution (разрешение кореференций), language detection (определение языка), sentiment analysis (анализ тональности), intent recognition (распознавание интенгов)
- **на уровне документов:** text summarization (автоматическое реферирование текста), machine translation (машинный перевод), topic modeling (тематическое моделирование)

Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)
- **на уровне слов:** error correction (исправление опечаток), normalization (нормализация: лемматизация или стемминг), morphological analysis (морфологический анализ)
- **на уровне словосочетаний:** POS-tagging (разметка частей речи), named entity recognition (извлечение именованных сущностей), word segmentation (сегментация слов)
- **на уровне предложений:** sentence segmentation (сегментация предложений), query answering (ответы на вопросы), word sense disambiguation (снятие лексической многозначности), syntax parsing (синтаксический разбор)
- **на уровне параграфов:** coreference resolution (разрешение кореференций), language detection (определение языка), sentiment analysis (анализ тональности), intent recognition (распознавание интенгов)
- **на уровне документов:** text summarization (автоматическое реферирование текста), machine translation (машинный перевод), topic modeling (тематическое моделирование)
- **на уровне корпуса:** deduplication (удаление дубликатов), informational retrieval (информационный поиск)

Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)
- **на уровне слов:** error correction (исправление опечаток), normalization (нормализация: лемматизация или стемминг), morphological analysis (морфологический анализ)
- **на уровне словосочетаний:** POS-tagging (разметка частей речи), **named entity recognition (извлечение именованных сущностей)**, word segmentation (сегментация слов)
- **на уровне предложений:** sentence segmentation (сегментация предложений), query answering (ответы на вопросы), word sense disambiguation (снятие лексической многозначности), syntax parsing (синтаксический разбор)
- **на уровне параграфов:** coreference resolution (разрешение кореференций), language detection (определение языка), sentiment analysis (анализ тональности), **intent recognition (распознавание интенгов)**
- **на уровне документов:** text summarization (автоматическое реферирование текста), machine translation (машинный перевод), topic modeling (тематическое моделирование)
- **на уровне корпуса:** deduplication (удаление дубликатов), informational retrieval (информационный поиск)

Основные задачи NLP

- **на уровне сигналов:** speech recognition (распознавание речи), speech synthesis (генерация речи)
- **на уровне слов:** error correction (исправление опечаток), normalization (нормализация: лемматизация или стемминг), morphological analysis (морфологический анализ)
- **на уровне словосочетаний:** POS-tagging (разметка частей речи), named entity recognition (извлечение именованных сущностей), word segmentation (сегментация слов)
- **на уровне предложений:** sentence segmentation (сегментация предложений), query answering (ответы на вопросы), word sense disambiguation (снятие лексической многозначности), syntax parsing (синтаксический разбор)
- **на уровне параграфов:** coreference resolution (разрешение кореференций), language detection (определение языка), sentiment analysis (анализ тональности), intent recognition (распознавание интенгов)
- **на уровне документов:** text summarization (автоматическое реферирование текста), machine translation (машинный перевод), topic modeling (тематическое моделирование)
- **на уровне корпуса:** deduplication (удаление дубликатов), informational retrieval (информационный поиск)

Передышка!

и время посмотреть чатик на наличие вопросов...



IR + NER

IR (intent recognition, распознавание интенгов) - часть NLU-модуля диалоговой системы, который отвечает за "понимание" смысла во фразе.

NER (named entity recognition) - тоже часть NLU, но уже умеет вычленять из фразы определенные слова и словосочетания, объединенные определенными свойствами.



IR + NER

IR (intent recognition, распознавание интенгов) - часть NLU-модуля диалоговой системы, который отвечает за "понимание" смысла во фразе.

NER (named entity recognition) - тоже часть NLU, но уже умеет вычленять из фразы определенные слова и словосочетания, объединенные определенными свойствами

IR + NER - основные ребята, которые отвечают за понимание смысла во фразе.

Причем формат этого понимания отличается от системы к системе. Но сама архитектура *"фраза отдается на вход некоторому черному ящику и из этого ящика на выходе получается некоторое представление смысла сказанного в формате, который умеет понимать диалоговый менеджер"* остается неизменной.

Распознавание интенгов и именованных сущностей на практике

>>> добрый день, подскажите, пожалуйста, какая процентная ставка у кредита "цветочный"?

Интегт: процентная_ставка_кредит

Сущности: "цветочный"

Интегт: ставка

Сущности: кредит цветочный

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**

Про датасеты и корпусы

Корпус - это набор текстов: больших, маленьких, по предложению, по параграфу, по слову...

Сложности:

- лицензия
- язык
- формат
- домен
- разметка
- размер

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация

```
from nltk import word_tokenize  
word_tokenize("Санкт-Петербург - северная столица России.")  
  
Out[8]: ['Санкт-Петербург', '-', 'северная', 'столица', 'России', '.']
```

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF
 - word2vec (GloVe, fasttext)

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF
 - word2vec (GloVe, fasttext)
 - BERT

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF
 - word2vec (GloVe, fasttext)
 - BERT
 - ... и все что вашей душеньке и модели угодно! (2)

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF
 - word2vec (GloVe, fasttext)
 - BERT
 - ... и все что вашей душеньке и модели угодно! (2)
- **модель**

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF
 - word2vec (GloVe, fasttext)
 - BERT
 - ... и все что вашей душеньке и модели угодно! (2)
- **модель**
 - classical ML

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF
 - word2vec (GloVe, fasttext)
 - BERT
 - ... и все что вашей душеньке и модели угодно! (2)
- **модель**
 - classical ML
 - deep learning

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF
 - word2vec (GloVe, fasttext)
 - BERT
 - ... и все что вашей душеньке и модели угодно! (2)
- **модель**
 - classical ML
 - deep learning
 - heuristics/patterns

Итак, классический процесс распознавания интенгов можно разделить на следующие шаги:

- **поиск данных**
- **подготовка данных**
 - токенизация
 - лемматизация
 - стемминг
 - исправление опечаток
 - удаление пунктуации
 - удаление стоп-слов
 - lowercase
 - ... и все что вашей душеньке и модели угодно!
- **выделение признаков/векторизация**
 - n-grams + TF-IDF
 - word2vec (GloVe, fasttext)
 - BERT
 - ... и все что вашей душеньке и модели угодно! (2)
- **модель**
 - classical ML
 - deep learning
 - heuristics/patterns
 - ... и снова все что вашей душеньке и модели угодно! (3)

На деле может происходить разное!

- rule-based

* you * [most] (amazing/gorgeous/fantastic)

*

На деле может происходить разное!

- rule-based

* you * [most] (amazing/gorgeous/fantastic)

*

- semantic-similarity/ranking

вопрос -> *ответ*

На деле может происходить разное!

- rule-based

* you * [most] (amazing/gorgeous/fantastic)

- semantic-similarity/ranking

вопрос -> *ответ*

- традиционная классификация

```
data.head()
```

| | text | class |
|---|--|----------|
| 0 | you look fantastic, actually | gorgeous |
| 1 | you're the most amazing creature I've ever met | gorgeous |
| 2 | honey, you're gorgeous! | gorgeous |
| 3 | you very smart i like it | smart |
| 4 | nice to meet clever people like you | smart |

На деле может происходить разное!

- rule-based
 - * you * [most] (amazing/gorgeous/fantastic)
- * semantic-similarity/ranking
 - *вопрос* -> *от*
- традиционная классификация
 - *time t





Но это ещё не всё!

Распознавание интенгов в NLU-модуле - не обязательно состоит из одной модели. Часто это пайплайн, или что-то вроде ансамбля (отличается от реализации к реализации).



Но это ещё не всё!

Распознавание интенгов в NLU-модуле - не обязательно состоит из одной модели. Часто это пайплайн, или что-то вроде ансамбля (отличается от реализации к реализации).

на “**крупнозерные**” тематики можно вешать тяжелые модели
на “**мелкозерные**” что-то узкое типа регулярок и шаблонов

Сделаем 5-секундную передышку

и перейдем к NER





NER

Мало понять, что вообще происходит в реплике пользователя, надо ещё и понять, о чем конкретно речь

- заказать пиццу **КАКУЮ?**
- отправить **СКОЛЬКО** рублей на **КАКОЙ** номер телефона?
- включить песню **КАКОГО** исполнителя?



NER

именованная сущность - набор слов, объединенный общей характеристикой



NER

именованная сущность - набор слов, объединенный общей характеристикой

неформально можно разделить на

- **общеупотребимые** (номер телефона, город)

NER

именованная сущность - набор слов, объединенный общей характеристикой

неформально можно разделить на

- общеупотребимые (номер телефона, город)
- **доменные** (соусы к картошке, названия кредитов в банке и тд)

NER может решаться



NER может решаться

- словарями (города, к примеру)

NER может решаться

- словарями (города, к примеру)
- грамматиками



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками
- ML (CRF, LSTM ...)



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками
- ML (CRF, LSTM ...)

В ЖИЗНИ:



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками
- ML (CRF, LSTM ...)

В жизни:

- `yargu`



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками
- ML (CRF, LSTM ...)

В жизни:

- yargy
- Natasha



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками
- ML (CRF, LSTM ...)

В жизни:

- yargy
- Natasha
- pymorphy



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками
- ML (CRF, LSTM ...)

В жизни:

- yargy
- Natasha
- pymorphy
- mystem



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками
- ML (CRF, LSTM ...)

В жизни:

- yargy
- Natasha
- pymorphy
- mystem
- deeppavlov



NER может решаться

- словарями (города, к примеру)
- грамматиками
- регулярками
- ML (CRF, LSTM ...)

В жизни:

- yargy
- Natasha
- pymorphy
- mystem
- deeppavlov
- ...

Всё это в теории, конечно, хорошо. Но есть жестокая жизнь, которая, иногда, не имеет ничего общего с теорией.

итак, ложка дегтя №1

в голосовых помощниках приходится мириться (или бороться) с
нежелательными артефактами от ASR

в голосовых помощниках приходится мириться (или бороться) с нежелательными артефактами от ASR

- ошибки распознавания

в голосовых помощниках приходится мириться (или бороться) с нежелательными артефактами от ASR

- ошибки распознавания
- “смешанная” речь (одновременно говорят другие)

в голосовых помощниках приходится мириться (или бороться) с нежелательными артефактами от ASR

- ошибки распознавания
- “смешанная” речь (одновременно говорят другие)
- промахнулись на этапе детекции начала реплики и/или её конца

в голосовых помощниках приходится мириться (или бороться) с нежелательными артефактами от ASR

- ошибки распознавания
- “смешанная” речь (одновременно говорят другие)
- промахнулись на этапе детекции начала реплики и/или её конца
- ...

ложка дегтя №2

Особенности и ограничения диалоговых систем, напрямую относящихся к NLU части

Особенности и ограничения диалоговых систем, напрямую относящихся к NLU части

- установка в контур заказчика

Особенности и ограничения диалоговых систем, напрямую относящихся к NLU части

- установка в контур заказчика
 - чувствительные данные
 - сопровождение
 - конфигурация системы

Особенности и ограничения диалоговых систем, напрямую относящихся к NLU части

- установка в контур заказчика
 - чувствительные данные
 - сопровождение
 - конфигурация системы
- время отклика

Особенности и ограничения диалоговых систем, напрямую относящихся к NLU части

- установка в контур заказчика
 - чувствительные данные
 - сопровождение
 - конфигурация системы
- время отклика
 - ГА - почти мгновенно
 - суфлеры/техсап - задержка не так критична

Особенности и ограничения диалоговых систем, напрямую относящихся к NLU части

- установка в контур заказчика
 - чувствительные данные
 - сопровождение
 - конфигурация системы
- время отклика
 - ГА - почти мгновенно
 - суфлеры/техсап - задержка не так критична
- количество данных на входе от заказчика

Особенности и ограничения диалоговых систем, напрямую относящихся к NLU части

- установка в контур заказчика
 - чувствительные данные
 - сопровождение
 - конфигурация системы
- время отклика
 - ГА - почти мгновенно
 - суфлеры/техсап - задержка не так критична
- количество данных на входе от заказчика
 - есть ли они вообще?
 - на что они пригодны?

(почти) ложка дегтя №3

NLU-модели имеют свойство “протухать”, поэтому сделать и забыть скорее всего не получится.

слайд для мотивации



*в детстве я мечтал стать
пожарным, но судьба что-то
перепутала и теперь я не тушу
а зажигаю*

Итоги!

NLU - чтобы наш ГА не сел в лужу и понял, что хотел сказать пользователь, иначе зачем это всё...

Итоги!

NLU - чтобы наш ГА не сел в лужу и понял, что хотел сказать пользователь, иначе зачем это всё...

Может быть разной степени изощренности:

- на регуляряках (no shame!)
- на TF-IDF + logreg
- на BERT'е и N-best
- пайплайн из n моделей
- вплоть до соответствия морфологических признаков
- или крутых навороченных DL моделек!



Итоги!

NLU - чтобы наш ГА не сел в лужу и понял, что хотел сказать пользователь, иначе зачем это всё...

Может быть разной степени изощренности:

- на регулярках (**no shame!**)
- на TF-IDF + logreg
- на BERT'е и N-best
- пайплайн из n моделей
- вплоть до соответствия морфологических признаков
- или крутых навороченных DL моделек!

Важно понимать NLU - это не только про эфемерное “о чем”, но и другая информация, полезная для дальнейшей работы бота.

Спасибо!

Засим откланиваюсь.

When you penalize your Natural Language Generation model for large sentence lengths



(мой любимый мем по NLP тематике)