

СТАРТАПУ ИИ: Технологии чатботов

Обсуждаем: <https://t.me/devdvAI>, <https://t.me/devdvStartup>

Репозиторий: <https://github.com/akumidv/startup-khv-ai-study>

ТИПЫ ЧАТБОТОВ

- GOAL-ORIENTED целеориентированные или для решения определенных задач
- ОТКРЫТЫЕ ЧАТБОТЫ «болталки»

DEEP LEARNING ЧАТБОТЫ

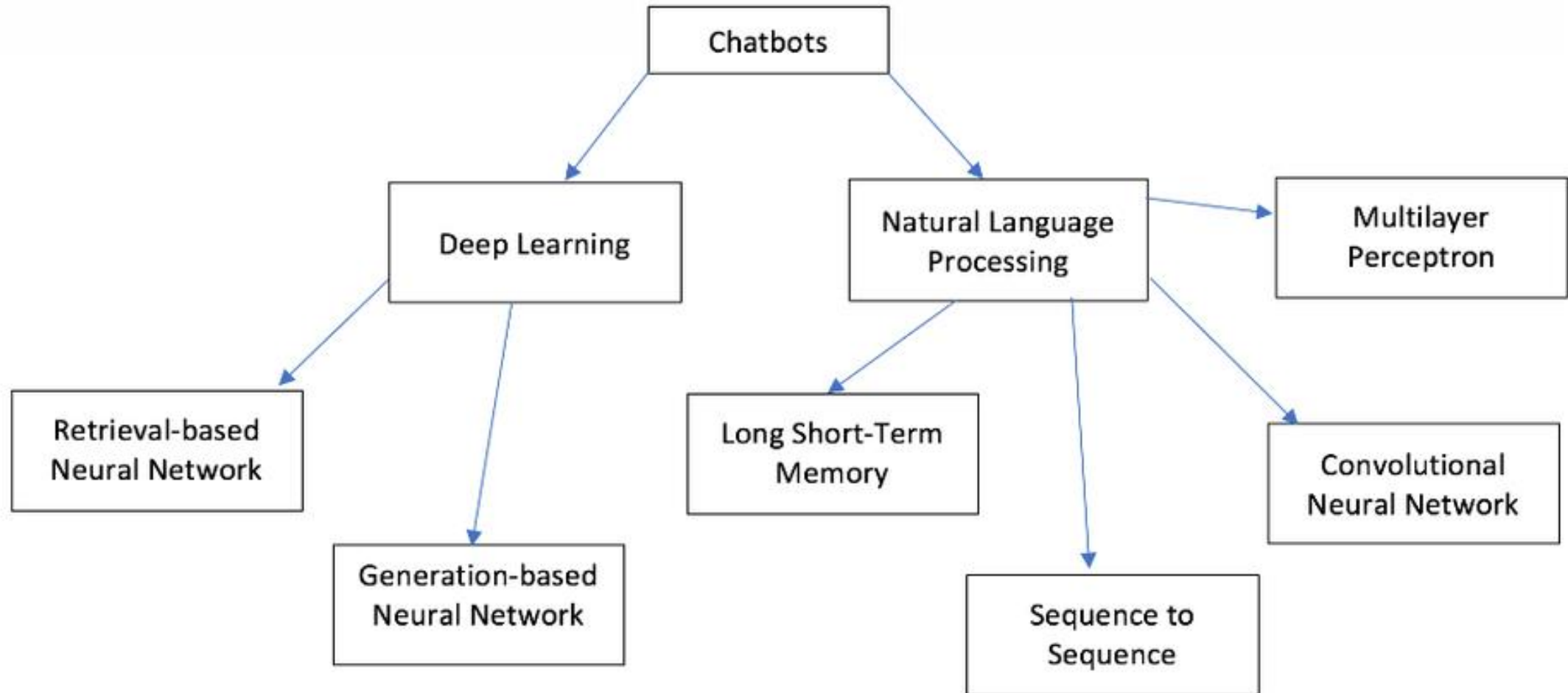
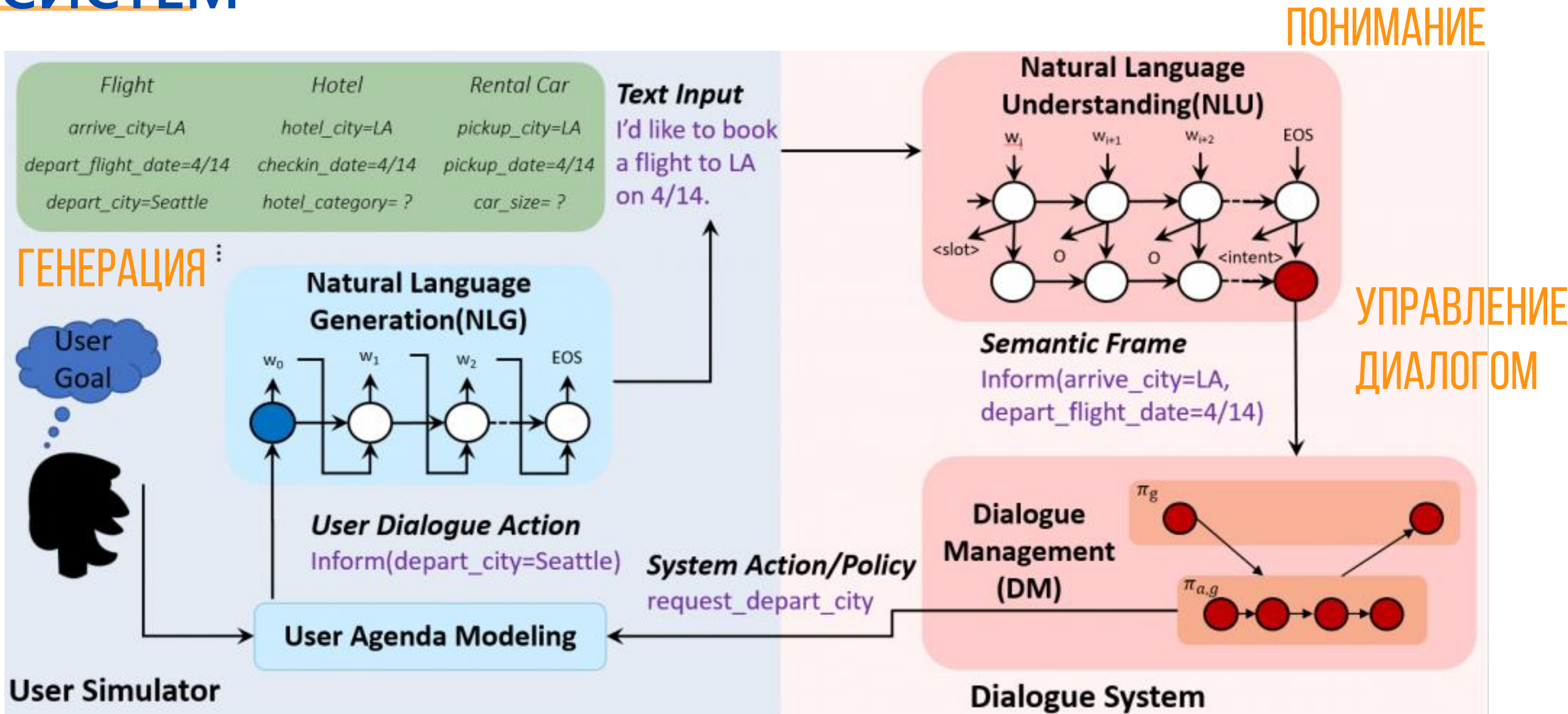


СХЕМА ЦЕЛЕОРИЕНТИРОВАННОГО ДИАЛОГОВЫХ СИСТЕМ



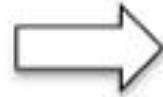
NLU: ПОНИМАНИЕ ЕСТЕСТВЕННОГО ЯЗЫКА

- Определяется:
 - домен (область знаний)
 - намерение(класс), желание пользователя во фразе
 - именованные сущности – слова отнесенные к типам

INTENT (НАМЕРЕНИЕ)

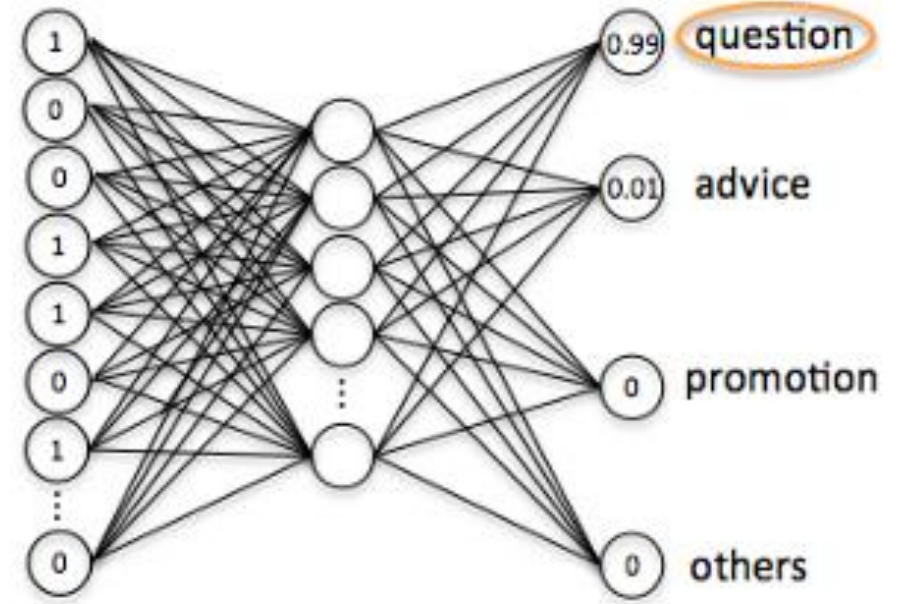
Mau tanya master.. Pupuk organik cair apa yang n lengkap menurut pengalaman teman2.?

Input text



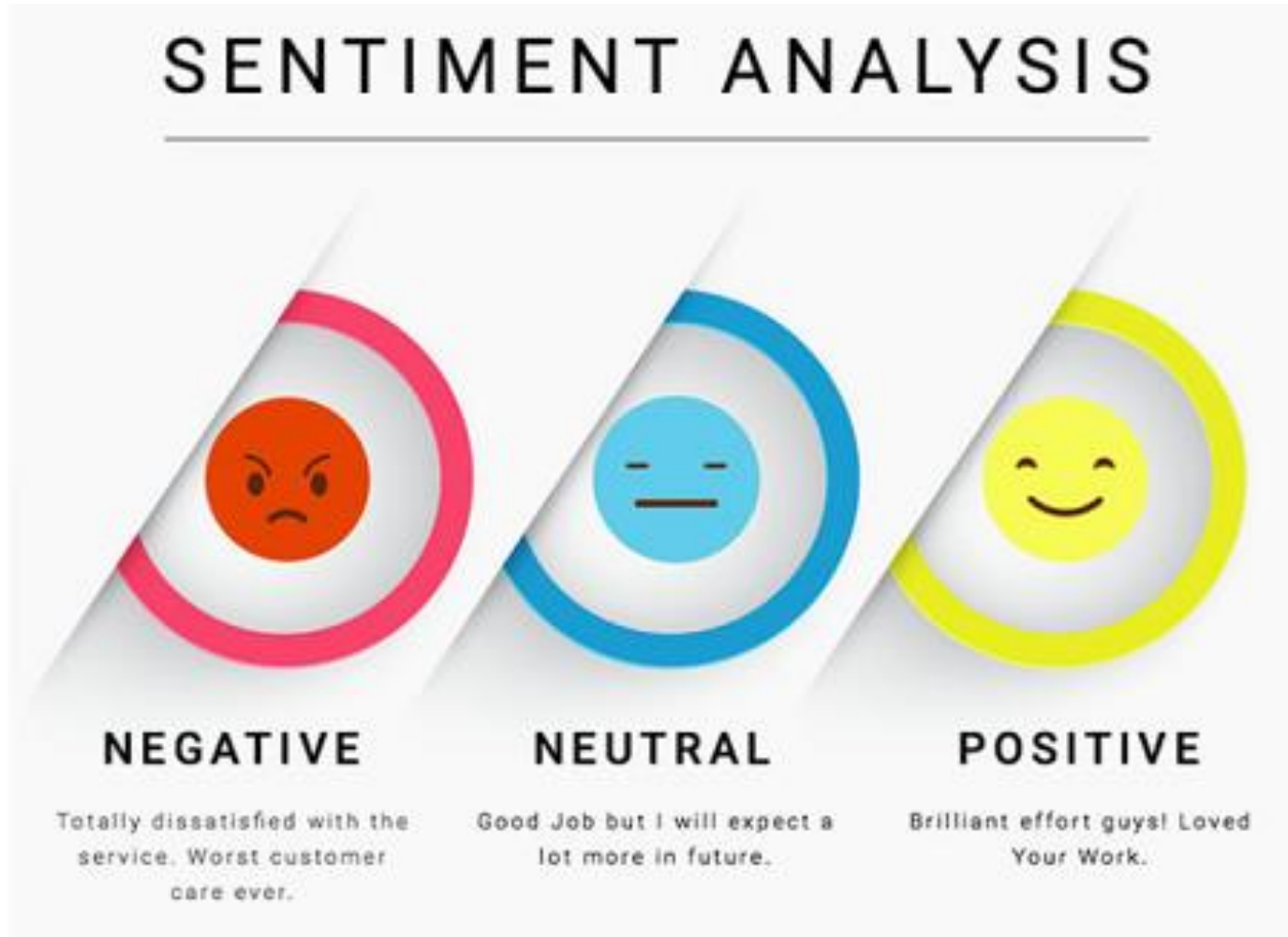
pupuk	1
bercak	0
padi	0
pengalaman	1
apa	1
air	0
lengkap	1
⋮	⋮
pestisida	0

**Word Embedding
Result**



**Neural Network
Classification**

ТОНАЛЬНОСТЬ (SENTIMENT)



[https://github.com/akumidv/startup-khv-ai-study/tree/main/06 Chatbots/0 technology/twitter_sentiment_ru_dataset_prepare.ipynb](https://github.com/akumidv/startup-khv-ai-study/tree/main/06%20Chatbots/0%20technology/twitter_sentiment_ru_dataset_prepare.ipynb)

NER: РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ

Российское издательство "Эксмо" с июня текущего года управляет издательским и журнальным бизнесом своего конкурента АСТ по договоренности с его акционерами . Об этом рассказал владелец и генеральный директор "Эксмо" Олег Новиков , пишет в номере от 26 июня газета "Ведомости" . Как указывает издание , "Эксмо" получило

Российское издательство "Эксмо" с июня тек...

Италия разместила на открытом рынке бонды ...

Спросить

Результаты

Российское издательство " **ЭКСМО** **ORG** " с июня текущего года управляет издательским и журнальным бизнесом своего конкурента **АСТ** **ORG** по договоренности с его акционерами . Об этом рассказал владелец и генеральный директор " **ЭКСМО** **ORG** " **Олег Новиков** **PER** , пишет в номере от 26 июня газета " **Ведомости** **ORG** " . Как указывает издание , " **ЭКСМО** **ORG** " получило трехлетний опцион на стопроцентный контроль в трех десятках юридических лиц группы **АСТ** **ORG** , в том числе в издательствах " **Астрель** **ORG** " , **АСТ** **ORG** , выпускающем журналы издательстве " **Премьера** **ORG** " и дистрибуторской компании " **Билония** **ORG** " . По словам **Новикова** **PER** , в сделку не входит сеть магазинов " **Буква** **ORG** " , а также ее недвижимость . По расчетам гендиректора " **ЭКСМО** **ORG** " , опцион может быть исполнен в течение года . **Новиков** **PER** отметил , что в ближайшее время " **ЭКСМО** **ORG** " намерено инвестировать 10 - 15 миллионов долларов в издательства и дистрибуцию **АСТ** **ORG** . Кроме того , владельцы **АСТ** **ORG** , хотя и не участвуют больше в управлении группой , имеют право на получение дивидендов в следующем году , до реализации опциона . С учетом дивидендов после осуществления опциона за весь бизнес владельцы **АСТ** **ORG** смогут получить около 70 миллионов долларов . В середине мая текущего года газета " **Коммерсантъ** **ORG** " со ссылкой на неназванные источники написала , что "

- Дистанция.
- Сумма денег
- Продолжительность времени
- Адреса электронной почты
- Числа [20ть человек]
- Порядковые числа [первый в очереди]
- Количество (вес)
- Время
- Интернет адрес
- Географический адрес
- Имена
- Номер телефона

УПРАВЛЕНИЕ ДИАЛОГОМ



TEXT QA

- Ответы на вопросы по тексту (Text QA)

Введите текст

Ньютон открыл всемирный закон тяготения Он сидел сидел под деревом, был сезон сбора фруктов. Яблоко упало ему на голову и у него родилась идея.

Введите вопрос

Как Ньютон открыл закон тяготения

Спросить

Результаты

О: Он сидел сидел под деревом

В: Как Ньютон открыл закон тяготения

Ньютон открыл всемирный закон тяготения Он сидел сидел под деревом **A** , был голову и у него родилась идея.

ODQA

- Open Domain Question Answering (ODQA) поиска ответа на любой вопрос внутри коллекции документов

Введите вопрос

Сколько жителей в Хабаровске|

Спросить

О: 1294

В: Сколько жителей в Хабаровске

О: в центральной части Белграда

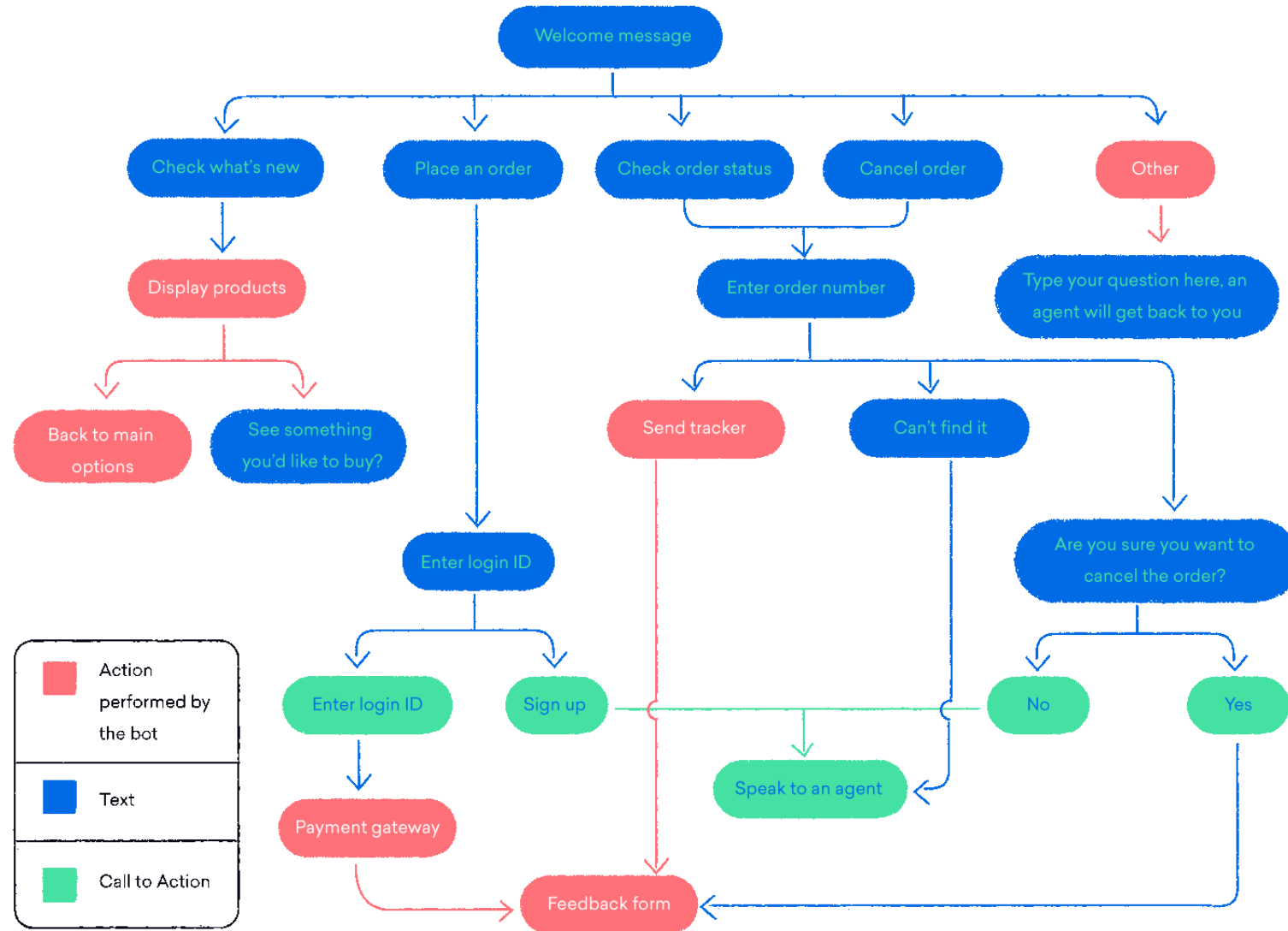
В: Где расположен международный аэропорт Никола Тесла?

ЧАТБОТЫ ПО ТИПАМ

НАКОНЕЦ СМОТРИМ КОД(!)

https://github.com/akumidv/startup-khv-ai-study/blob/main/06_Chatbots/

RULE – НА ОСНОВЕ ПРАВИЛ



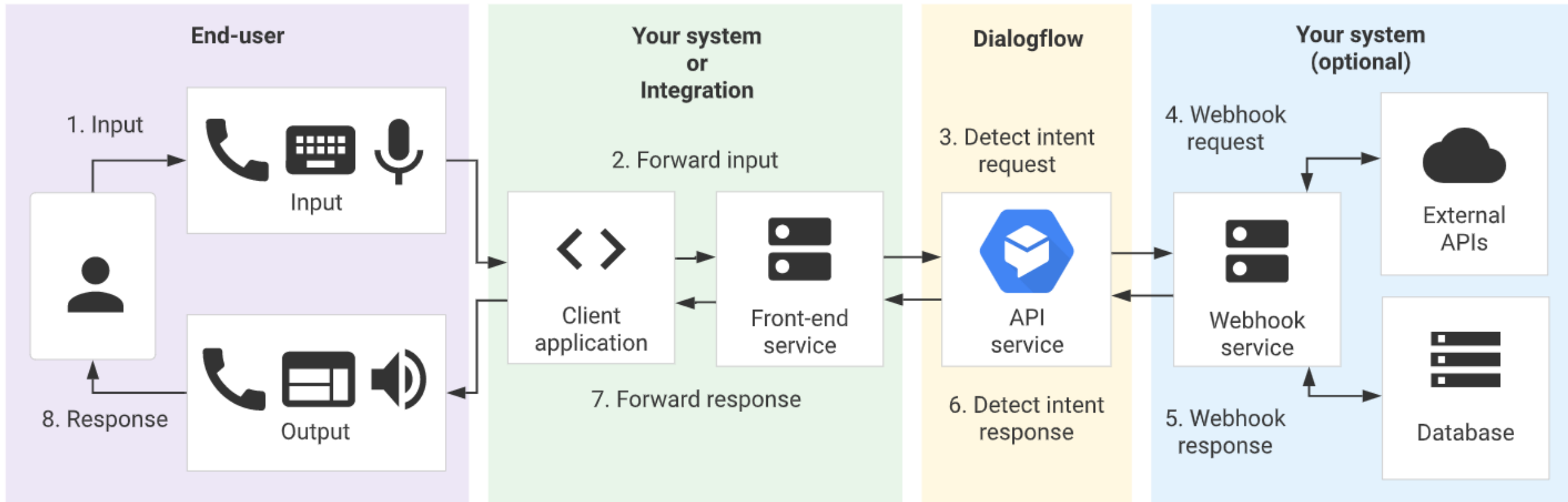
RETRIEVAL-BASED

РЕПОЗИТОРИЙ ГОТОВЫХ ОТВЕТОВ И ДЕЙСТВИЙ ИСПОЛЬЗУЕМЫХ В
СООТВЕТСТВИИ С ОПРЕДЕЛЁННЫМ КОНТЕКСТОМ

Контекст можно определять разными способами:

- Ключевые слова
- Вероятностная близость (баес)
- Близость векторов слов
- Нейросетевые модели



GOOGLE DIALOGFLOW



INTENTS: ДЕТЕКЦИЯ НАМЕРЕНИЙ


Intents


Search intents


-  Default Fallback Intent
-  Default Welcome Intent
-  Worry


Training phrases

 Add user expression

 привет привет

 здравствуй

 привет


 приветствую


 всем привет

 хей

 чао

 приветствую тебя

 приветик

 и снова здравствуйте


Responses

DEFAULT 

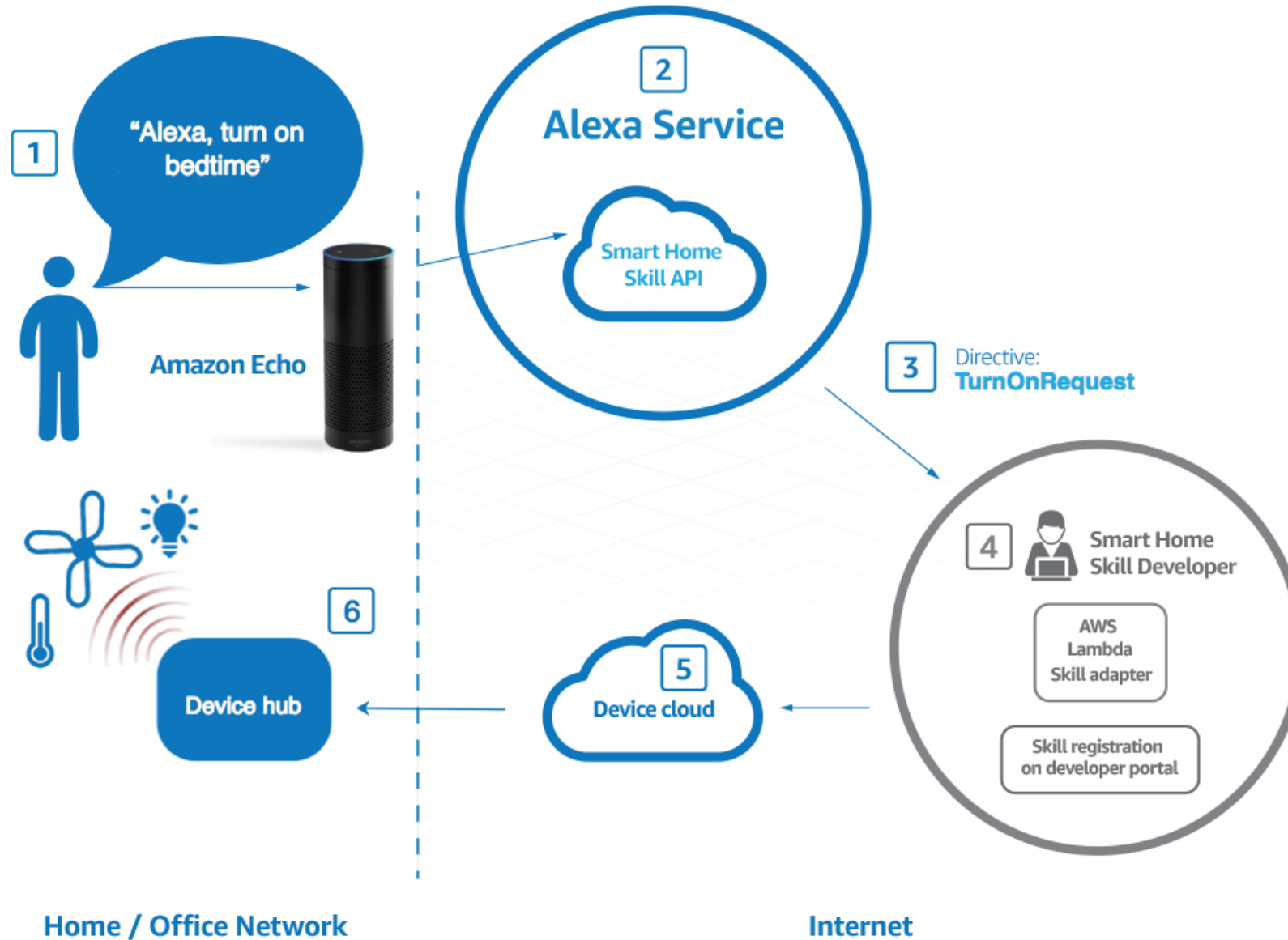
Text Response

- 1 Привет!
- 2 Здравствуй!
- 3 Добрый день!
- 4 Enter a text response variant

ADD RESPONSES

☐ Set this intent as end of conversation 

AMAZON ALEXA



AMAZON ALEXA: INTENTS AND CODE

Intents

+ Add Intent

NAME

AMAZON.FallbackIntent

AMAZON.StopIntent

AMAZON.CancelIntent

AMAZON.HelpIntent

AMAZON.NoIntent

AMAZON.YesIntent

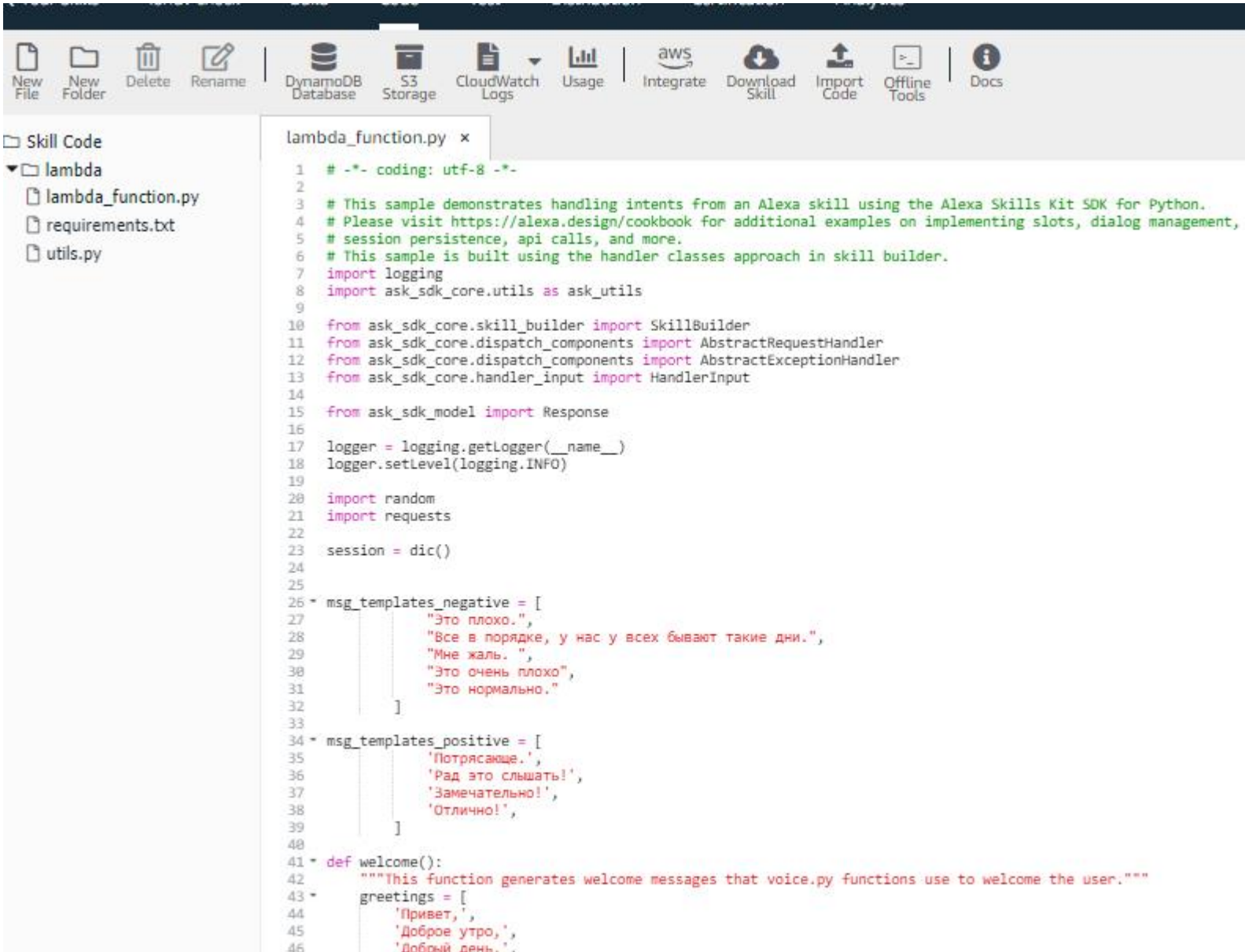
AMAZON.PreviousIntent

AMAZON.StartOverIntent

PositiveFeeling

NegativeFeeling

BedYes



The screenshot shows an IDE interface. On the left, a file explorer displays the 'Skill Code' directory containing a 'lambda' subdirectory with files 'lambda_function.py', 'requirements.txt', and 'utils.py'. The main editor window shows the 'lambda_function.py' file with the following Python code:

```
1  # -*- coding: utf-8 -*-
2
3  # This sample demonstrates handling intents from an Alexa skill using the Alexa Skills Kit SDK for Python.
4  # Please visit https://alexa.design/cookbook for additional examples on implementing slots, dialog management,
5  # session persistence, api calls, and more.
6  # This sample is built using the handler classes approach in skill builder.
7  import logging
8  import ask_sdk_core.utils as ask_utils
9
10 from ask_sdk_core.skill_builder import SkillBuilder
11 from ask_sdk_core.dispatch_components import AbstractRequestHandler
12 from ask_sdk_core.dispatch_components import AbstractExceptionHandler
13 from ask_sdk_core.handler_input import HandlerInput
14
15 from ask_sdk_model import Response
16
17 logger = logging.getLogger(__name__)
18 logger.setLevel(logging.INFO)
19
20 import random
21 import requests
22
23 session = dic()
24
25
26 msg_templates_negative = [
27     "Это плохо.",
28     "Все в порядке, у нас у всех бывают такие дни.",
29     "Мне жаль. ",
30     "Это очень плохо",
31     "Это нормально."
32 ]
33
34 msg_templates_positive = [
35     'Потрясающе.',
36     'Рад это слышать!',
37     'Замечательно!',
38     'Отлично!',
39 ]
40
41 def welcome():
42     """This function generates welcome messages that voice.py functions use to welcome the user."""
43     greetings = [
44         'Привет,',
45         'Доброе утро,',
46         'Добрый день.'
```

ALEXA: TEST

Skill testing is enabled in: Development

☒ Skill I/O ☒ Device Display ☐ Device Log

Alexa Simulator Manual JSON Voice & Tone

English (US) Type or click and hold the mic

open test chat bot

Welcome, you can say Hello or Help. Which would you like to try?

hello

Hello W Alexa Simulator

Hello World!

Skill Invocations | Viewing: 1 / 1

JSON Input 1

```
1 {
2   "version": "1.0",
3   "session": {
4     "new": false,
5     "sessionId": "amzn1.echo-api.session.47774e63-8a50-4492-a5de-c90801ed",
6     "application": {
7       "applicationId": "amzn1.ask.skill.c5e3eb69-b5d8-4750-b896-b58a278"
8     },
9     "user": {
10      "userId": "amzn1.ask.account.AGL2QTHMNWJ5C4XMAHJQWB5F2J4VBYVDZGUY"
11    }
12  },
13  "context": {
14    "Viewports": [
15      {
16        "type": "APL",
17        "id": "main",
18        "shape": "RECTANGLE",
19        "dpi": 213,
20        "presentationType": "STANDARD",
21        "canRotate": false,
22        "configuration": {
23          "current": {
24            "mode": "HUB",
25            "video": {
26              "codecs": [
27                "H_264_42",
28                "H_264_41"
29              ]
30            },
31            "size": {
32              "type": "DISCRETE",
33              "pixelWidth": 1280,
34              "pixelHeight": 800

```

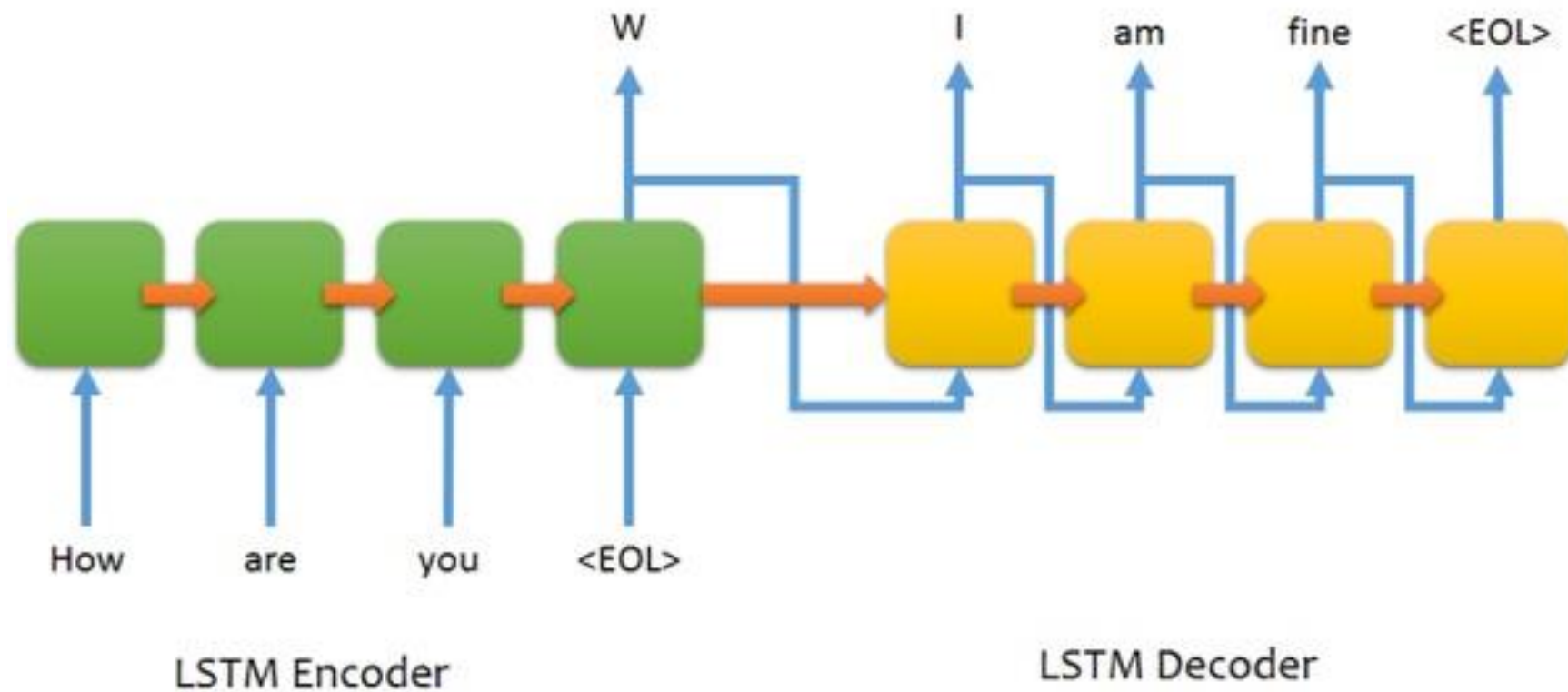
JSON Output 1

```
1 {
2   "body": {
3     "version": "1.0",
4     "response": {
5       "outputSpeech": {
6         "type": "SSML",
7         "ssml": "<speak>Hello World!</speak>"
8       },
9       "type": "_DEFAULT_RESPONSE"
10    },
11    "sessionAttributes": {},
12    "userAgent": "ask-python/1.11.0 Python/3.7.10"
13  }
14 }
```

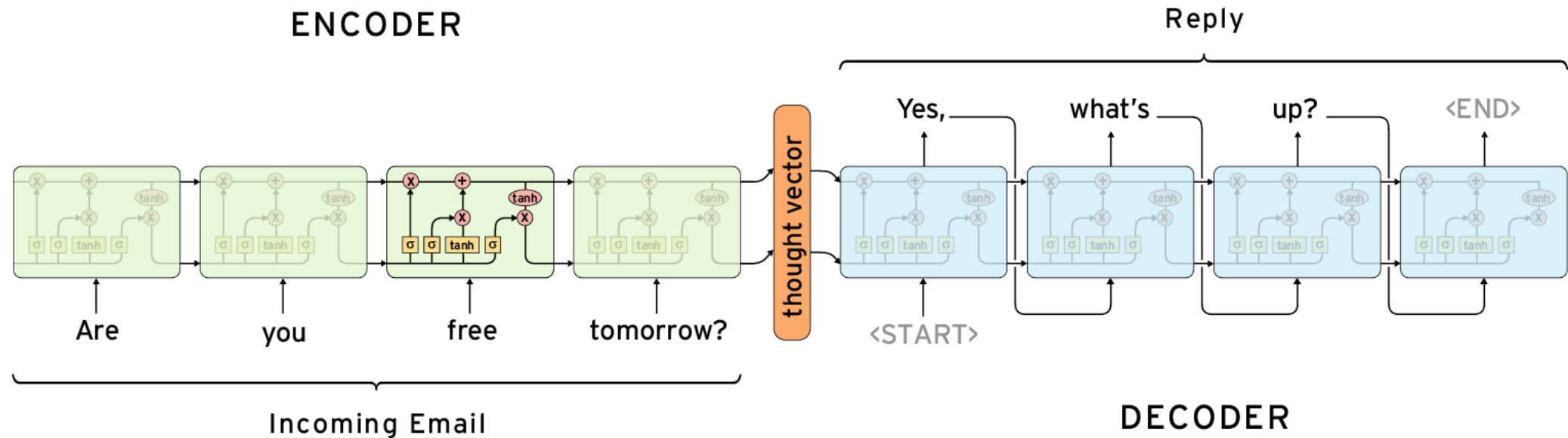
SEQ2SEQ – ЧАТБОТ КАК МАШИННЫЙ ПЕРЕВОД

ENCODER-DECODER

РЕПЛИКИ ПОЛЬЗОВАТЕЛЯ «ПЕРЕВОДЯТСЯ» В ПОДХОДЯЩИЕ ОТВЕТЫ НА НИХ



SEQ2SEQ: НЕДОСТАТКИ



- Общие фразы для любой ситуации:
 - I DON'T KNOW.
 - OK, I SEE
- Ответ из последней реплики, не помнит о чем речь раньше
 - Не помнит свои ответы
 - Не помнит вопросы/информацию пользователя

ГИБРИДЫ

ЧАТ БОТ ВИКА: ПРИМЕР ТЕХНОЛОГИЙ [HTTPS://GITHUB.COM/KOZIEV/CHATBOT](https://github.com/koziev/chatbot)

Гибридный подход – ML-based и правила. Некоторые виды правил генерируются из слабоструктурированных датасетов (chit-chat stories, continuation rules).

Retrieval-based + generative движок – ответы бота строятся на основе информации в базе знаний, обеспечивая консистентность ответов на перефразировки вопросов.

Динамика базы знаний – новые факты могут добавляться в базу по ходу диалога, обеспечивая боту долговременную память в рамках всей сессии и за ее пределами.

Проактивность – бот стремится продолжить диалог, задает пользователю вопросы для пополнения базы знаний, интерпретирует ответы на основе оперативного контекста.

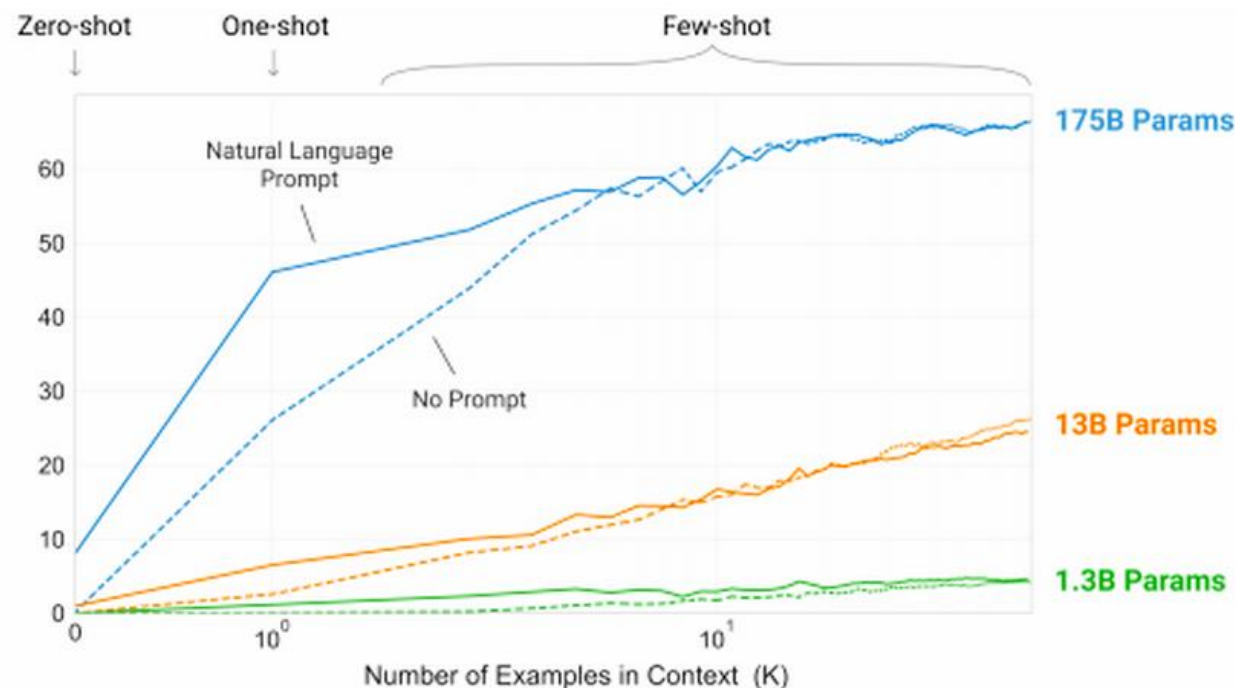
ГИБРИДЫ

Пример технологий <https://github.com/Koziev/chatbot> :

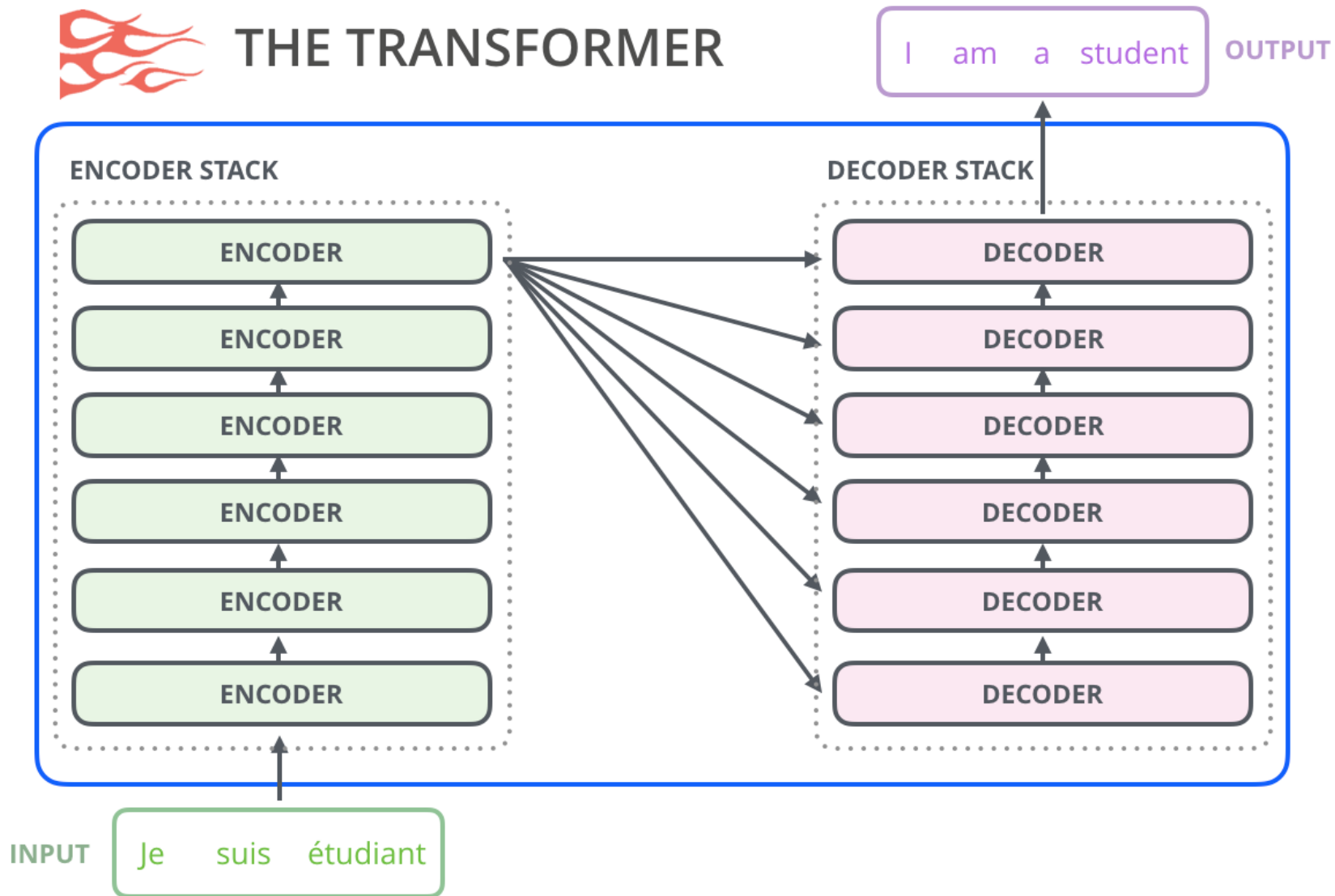
- Определение синонимии фраз [nn_synonymy_detector.py](#)
- Интерпретация реплики собеседника (раскрытие анафоры, эллипсиса, гэппинга, дополнение ответа etc) [nn_interpreter_new2.py](#)
- Определение релевантности предпосылки и вопроса [lgb_relevancy_detector.py](#)
- Генерация текста ответа с помощью seq2seq нейросетки [train_nn_seq2seq_pqa_generator.py](#)
- Посимвольное встраивание слово в вектор фиксированной длины [wordchar2vector_model.py](#)
- Определение достаточности набора предпосылок для генерации ответа [nn_enough_premises_model.py](#)
- NER для некоторых типов сущностей [entity_extractor.py](#)

GPT3: РАЗМЕР ИМЕЕТ ЗНАЧЕНИЕ

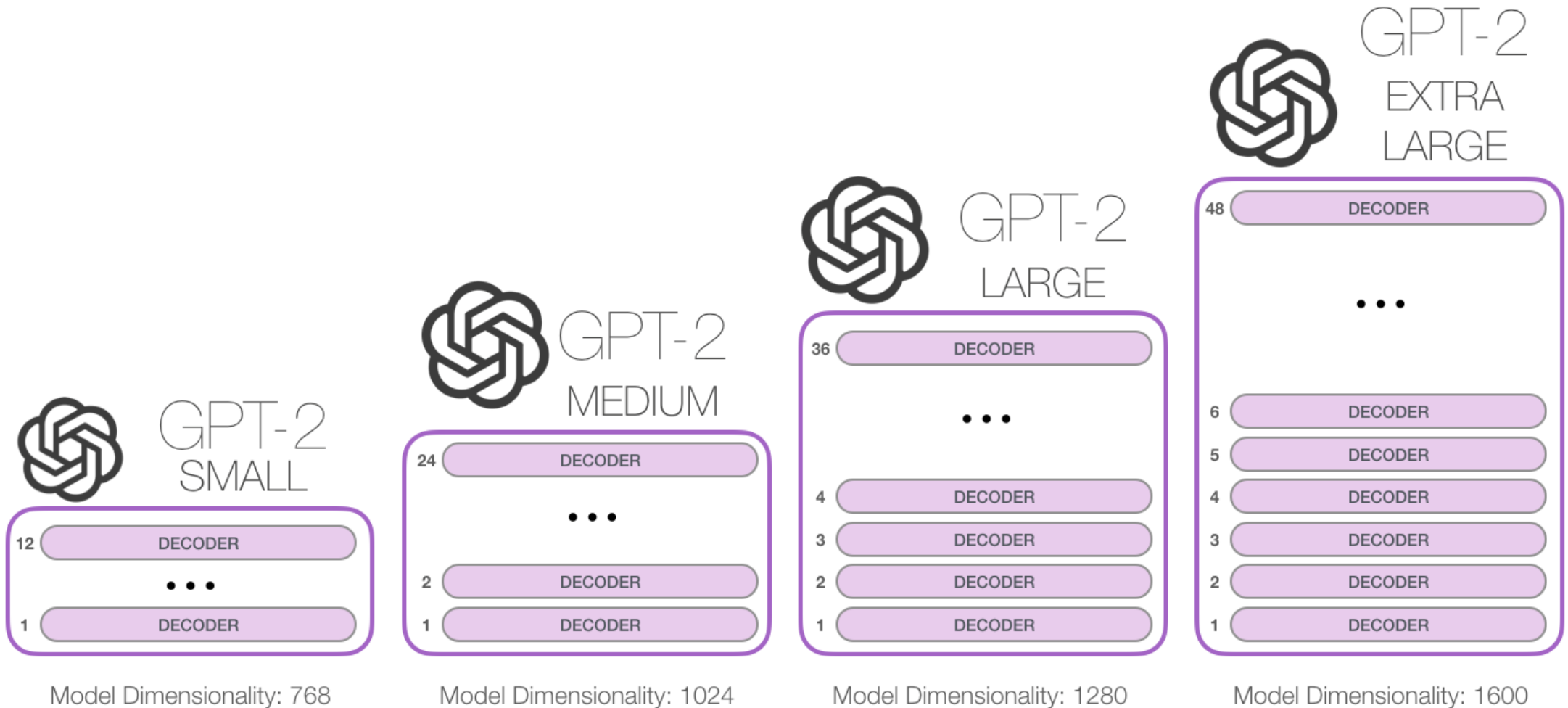
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}



АРХИТЕКТУРА GPT2



ОТЛИЧИЯ РАЗМЕРОВ В АРХИТЕКТУРЕ GPT2



ruGPT3

Model Name	nparams
ruGPT-3 Small	125M
ruGPT-3 Medium	350M
ruGPT-3 Large	760M
ruGPT-3 XL	1.3B
GPT-3 2.7B	2.7B
GPT-3 6.7B	6.7B
ruGPT-3 13B	13.0B
GPT-3 175B(GPT-3)	175.0B

КАЧЕСТВО GPT2 (DIALOGPT, 2019 Г.)

Method	NIST		BLEU		METEOR	Entropy E-4	Dist		Avg Len
	N-2	N-4	B-2	B-4			D-1	D-2	
PERSONALITYCHAT	0.19	0.20	10.44%	1.47%	5.42%	6.89	5.9%	16.4%	8.2
Team B	2.51	2.52	14.35%	1.83%	8.07%	9.03	10.9%	32.5%	15.1
Ours(117M)	1.58	1.60	10.36%	2.02%	7.17%	6.94	6.2%	18.94%	13.0
GPT(345M)	1.78	1.79	9.13%	1.06%	6.38%	9.72	11.9%	44.2%	14.7
Ours(345M)	2.80	2.82	14.16%	2.31%	8.51%	10.08	9.1%	39.7%	16.9
Ours(345M,Beam)	2.92	2.97	19.18%	6.05%	9.29%	9.57	15.7%	51.0%	14.2
Human	2.62	2.65	12.35%	3.13%	8.31%	10.45	16.7%	67.0%	18.8

TEAM B – ПОБЕДИТЕЛЬ 2018 DIALOG SYSTEM TECHNOLOGY CHALLENGE 7

КАК ФОРМИРОВАТЬ ДАТАСЕТ ДЛЯ ДИАЛОГОВ

|0|1|Привет, как дела?|1|-|Хорошо|0|1|Что нового?|1|3|

ПЕРВАЯ ЦИФРА 0 ИЛИ 1 – ВОПРОС И ОТВЕТ

ВТОРАЯ ЦИФРА 1,2,3 ИЛИ '-' - ДЛИНА

ВАРИАНТ DIALOGPT

```
0.0 what are you doing for a living ?<->1.0 i am a admin .^M
0.0 what are you doing for a living ?<->1.0 i am a admin .^M
0.0 what are you doing for a living ?<->1.0 i am a engineer .^M
0.0 what are you doing for a living ?<->1.0 i am a lawyer .^M
0.0 what is your favorite color ?<----->1.0 i love red .^M
0.0 what is your favorite color ?<----->1.0 i love red .^M
0.0 what is your favorite color ?<----->1.0 i love red .^M
```


КРИТЕРИИ КАЧЕСТВА ОБУЧЕНИЯ GPT3 DEV DV

- Исходный датасет 36 тыс. пар
- Перплексия:
 - при последовательном формировании пар сообщения 1+2, 3+4... на 5ти эпохах (6 тыс.шагов): `perplexity = tensor(35.8258)`
 - при нарастающей схеме 1+2, 2+3, 3+4
 - 5ть эпох `perplexity = tensor(17.2)`
 - 15ть эпох `perplexity = tensor(9.9298)`

Perplexity какова вероятность, что система сгенерирует правильный ответ (то есть ответ, который дал в этой ситуации пользователь). Или на тестовой выборке

ЧЕМ НИЖЕ, ТЕМ ЛУЧШЕ

DIALOGPT3 НА DEVDV И DEVDVAI КАНАЛАХ

Введите quit для остановки

User:Привет

Setting `pad_token_id` to `eos_token_id`:50257 for open-end generation.

==> debug Params generate: {'max_length': 256, 'no_repeat_ngram_size': 3, 'top_k': 10

==> debug Text input: |0|1|Привет |1|1|

>>>Bot: Есть ещё один интересный момент. Если предприниматель - это человек

all variants ==> debug: ['Есть ещё один интересный момент. Если предприниматель -

User:есть тут программисты

Setting `pad_token_id` to `eos_token_id`:50257 for open-end generation.

==> debug Params generate: {'max_length': 256, 'no_repeat_ngram_size': 3, 'top_k': 10

==> debug Text input: |0|1|Привет |1|1|Есть ещё один интересный момент. Если пре

>>>Bot: 2]В Хабаровске появится «Спортивный городок» <https://habarov.today/2019-04-28/>

all variants ==> debug: ['2]В Хабаровске появится «Спортивный городок» <https://habarov.today/2019-04-28/>

USER:Привет!

Setting `pad_token_id` to `eos_token_id`:50257 for open-end generation.

[debug] Затравка вопроса для модели: |0|1|Привет!|1|1|

>>>BOT: Ребзя, привет. Сегодня с утра иду знакомится с местными ИТ-компаниями. Им нужны ИТ-шни

[debug] other variants ==> : ['Всем привет! Давно не виделись, но уже столько всего с

RUGPT3 SMALL

36 ТЫС. СООБЩЕНИЙ

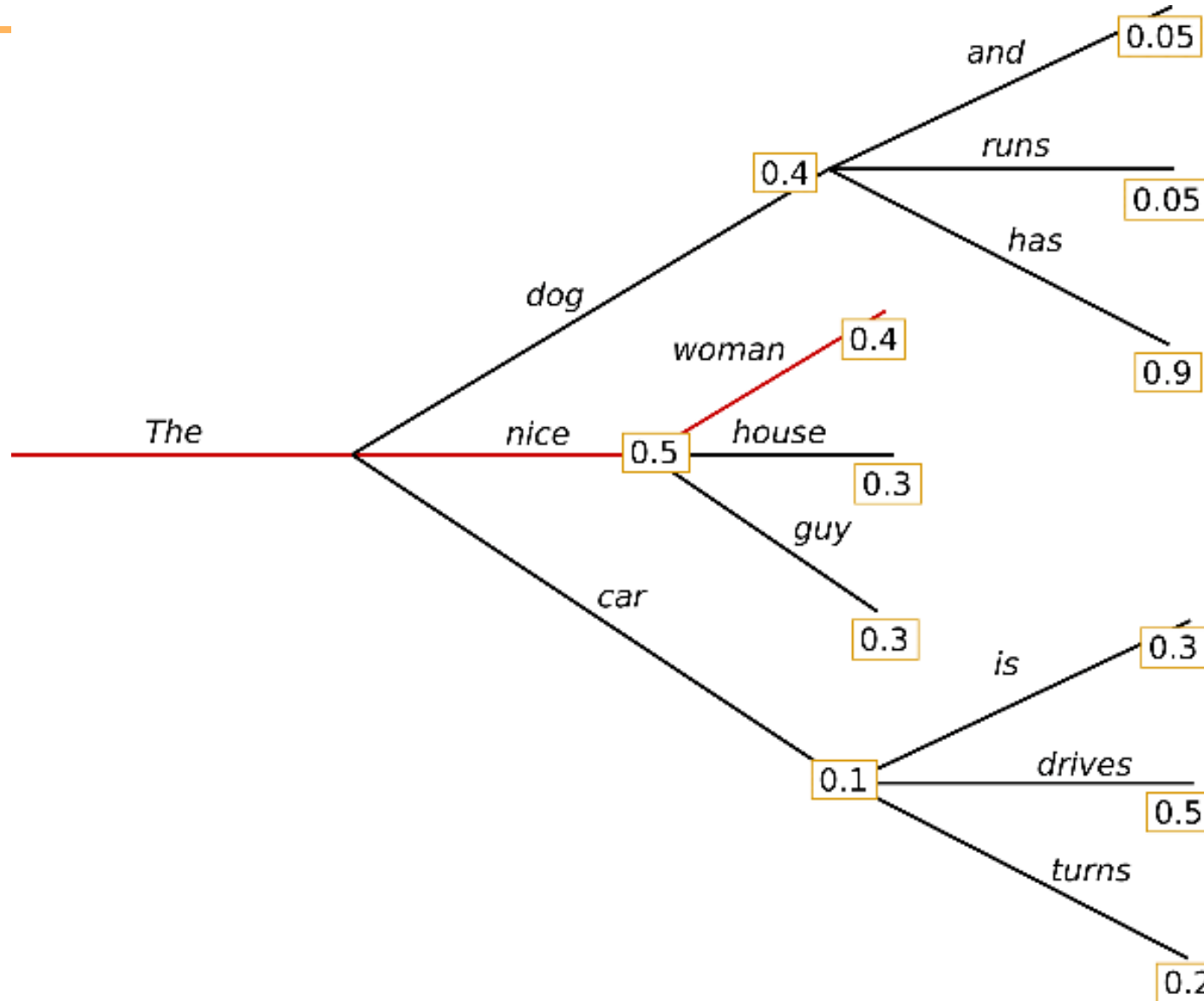
5 ЭПОХ,
ПЕРПЛЕКСИЯ 17ТЬ

15 ЭПОХ,
ПЕРПЛЕКСИЯ 9ТЬ

КАК GPT ВЫБИРАЕТ ОТВЕТ, ПАРАМЕТРЫ

<https://huggingface.co/blog/how-to-generate>

GREEDY SEARCH: СУММА ВЕРОЯТНОСТЕЙ

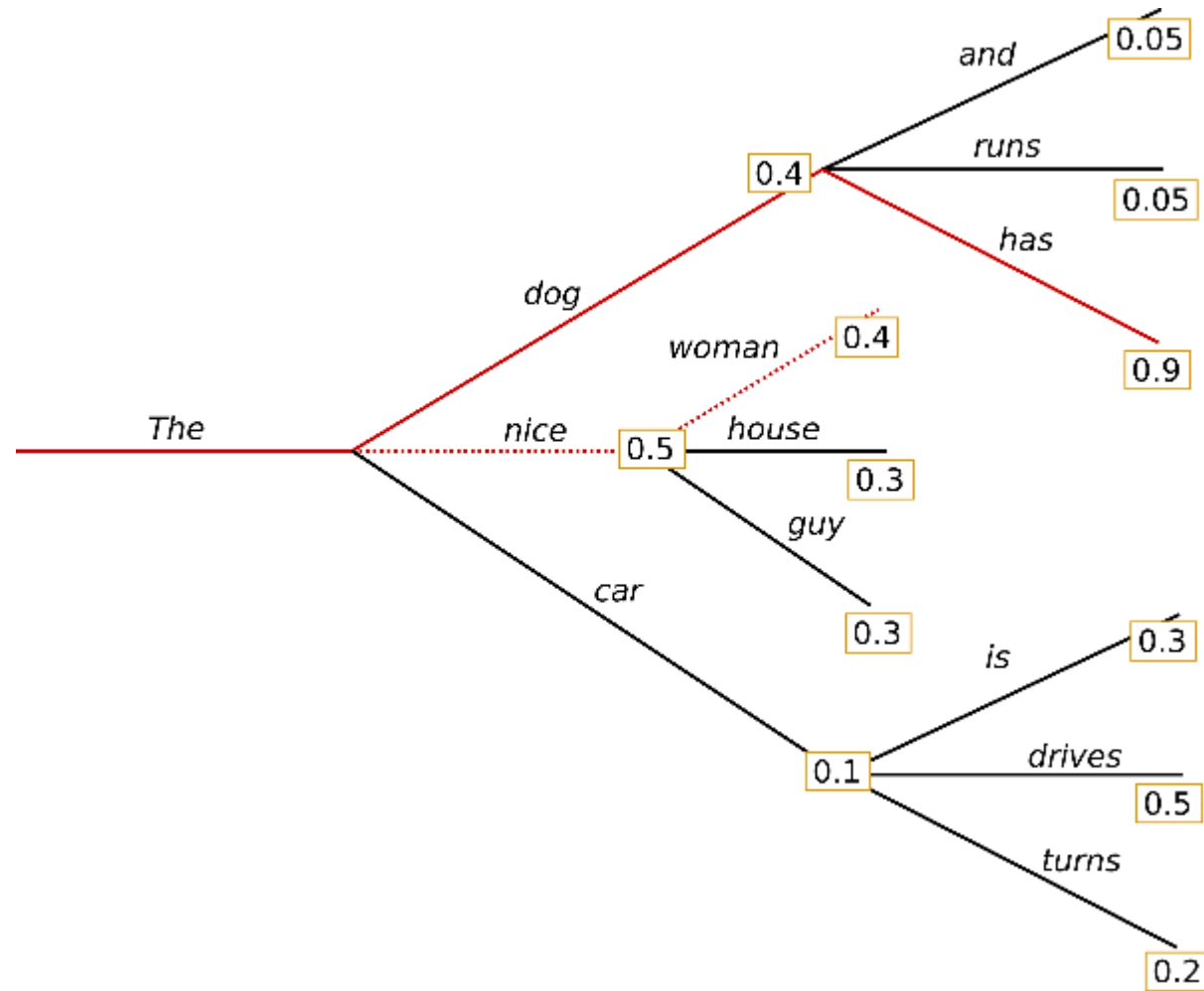


Вероятность:
 $0.5 \times 0.4 = 0.2$

I enjoy walking with my cute dog, but I'm not sure if

I'll ever be able to walk with my dog. I'm not sure if I'll

BEAM SEARCH: НАИБОЛЬШАЯ ВЕРОЯТНОСТЬ



Вероятность: 0.36

I enjoy walking with my cute dog, but I'm not sure if I'll

ever be able to walk with him again. I'm not sure if

BEAM SEARCH: НАИБОЛЬШАЯ ВЕРОЯТНОСТЬ

- **no_repeat_ngram_size** – сколько раз сочетание слов может встречаться
- I've been thinking about this for a while now, and I think it's time for me to take a break
- **NO New York** в тексте отобразится один раз... если указать 2.
- **num_return_sequences** - сколько вероятных ответов возвращать

I enjoy walking with my cute dog, but I'm not sure if I'll

ever be able to walk with him again. I'm not sure if

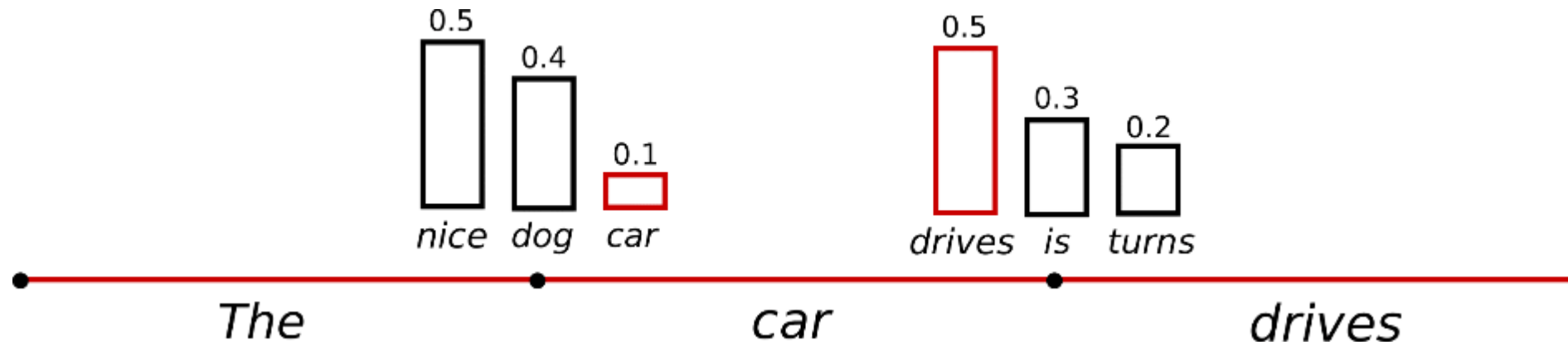
ПОВТОРЕНИЯ

- **no_repeat_ngram_size** – сколько раз сочетание слов может встречаться
- I've been thinking about this for a while now, and I think it's time for me to take a break
- **NO New York** в тексте отобразится один раз... если указать 2.
- **num_return_sequences** - сколько вероятных ответов возвращать

I enjoy walking with my cute dog, but I'm not sure if I'll

ever be able to walk with him again. I'm not sure if

SAMPLING - ВЫБОРКА

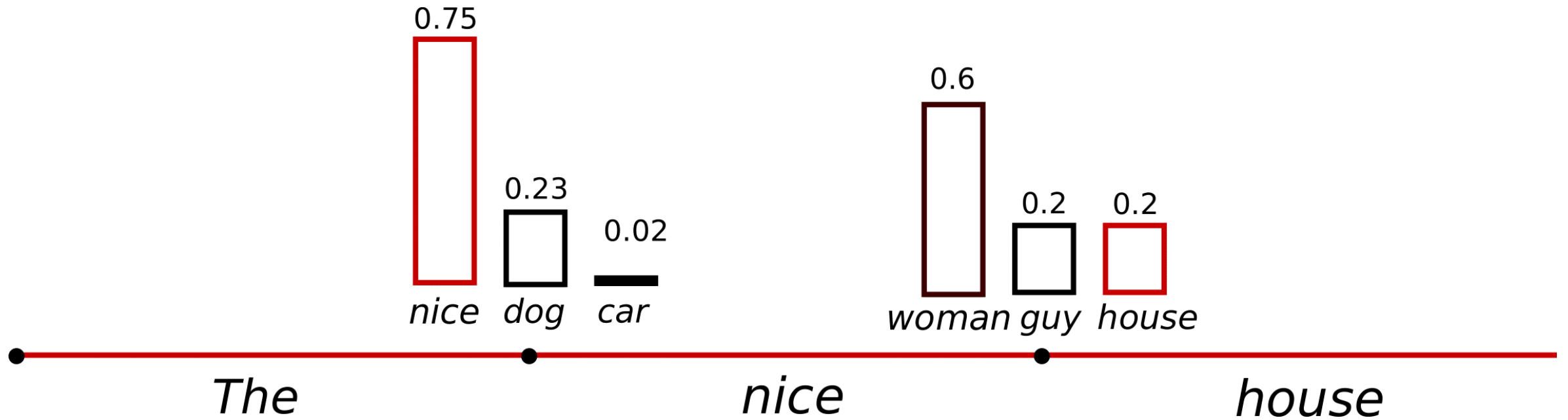


Выбираем не детерминированно, а по случайно по условному распределению вероятности

В библиотеке Transformers:

`do_sample=True`, при этом деактивируем *Top-K* sampling `top_k=0`

TEMPERATURE: ПАРАМЕТР SOFTMAX НЕЙРОНОВ



Делаем контрастней выбор: увеличиваем более вероятные, уменьшаем менее вероятные путем уменьшения температуры

TOP-K SAMPLING: ОГРАНИЧЕНИЕ КОЛ-ВА СЛОВ

1.0

ТОР-К 6ТЬ СЛОВ

$$\sum_{w \in V_{\text{top-K}}} P(w | \text{"The"}) = 0.68$$

0.0

nice dog car woman guy man people big house cat

$P(w | \text{"The"})$

$$\sum_{w \in V_{\text{top-K}}} P(w | \text{"The"}, \text{"car"}) = 0.99$$

drives is turns stops down a not the small told

$P(w | \text{"The"}, \text{"car"})$

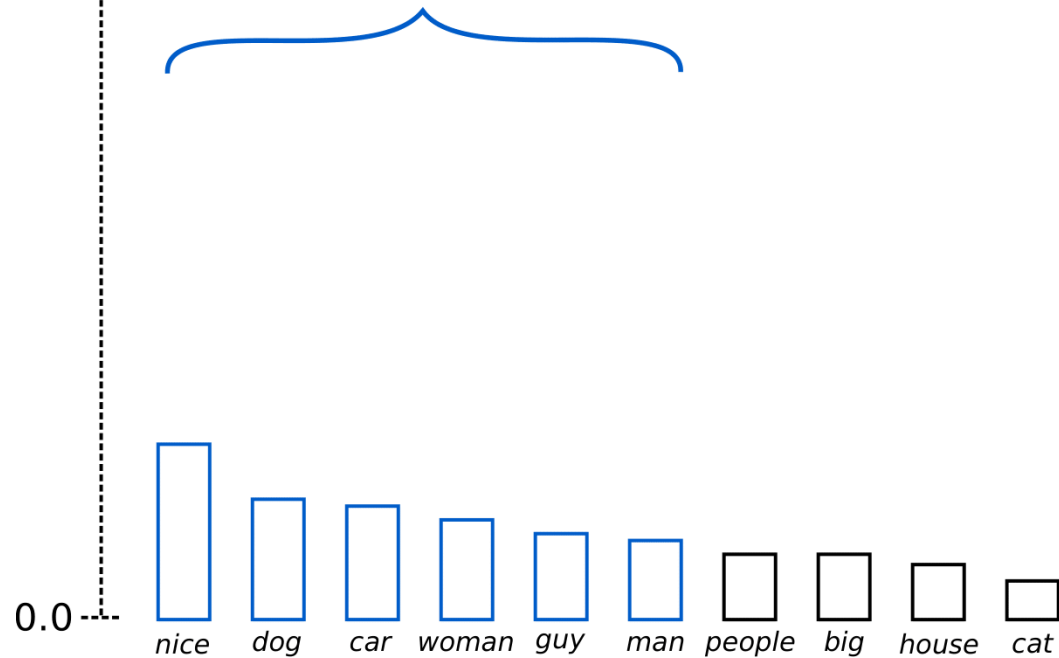
По добавлению слов в ответ, вероятность остальных падает

TOP-N SAMPLING: ЯДРО ВЫБОРКИ

1.0

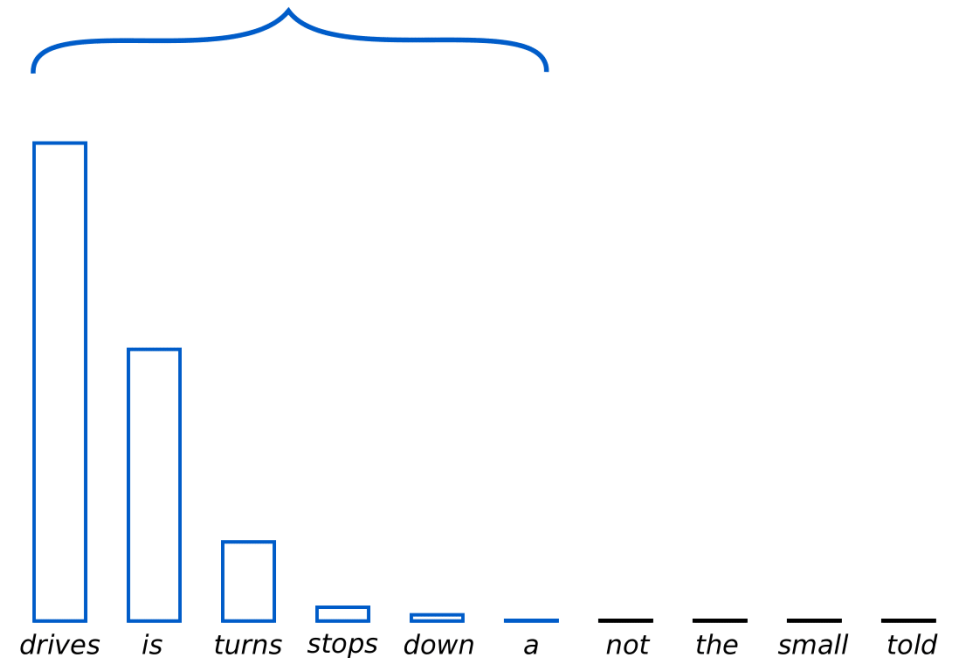
$P = 0,92$

$$\sum_{w \in V_{\text{top-K}}} P(w | \text{"The"}) = 0.68$$



$P(w | \text{"The"})$

$$\sum_{w \in V_{\text{top-K}}} P(w | \text{"The"}, \text{"car"}) = 0.99$$



$P(w | \text{"The"}, \text{"car"})$

Выбирает из наименьшего возможного набора слов, совокупная вероятность которых превышает вероятность p

КОНТАКТЫ

ОБСУЖДАЕМ

<https://t.me/devdvAI>

<https://t.me/devdvStartup>



РЕПОЗИТОРИЙ

<https://github.com/akumidv/startup-khv-ai-study>

АНДРЕЙ КУМИНОВ

+7 914 770 5846

<https://facebook.com/akuminov>

<https://vk.com/akumidv>