



Генерация текста

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна
Большакова Елена Игоревна

Содержание



- ☐ История и современное состояние
- ☐ Направления генерации текстов на естественном языке (ГЕЯ)
- ☐ Методы
 - ☐ Шаблоны
- ☐ Генерация правдоподобных текстов
 - ☐ Виртуальные собеседники
- ☐ Системы, генерирующие разумный текст
 - ☐ История. Теория и разработки
 - ☐ Этапы (модули) генерации
 - ☐ Примеры систем

История и современное состояние



- Первые текстовые сообщения – ответы операционной системы на естественном языке («done»)
- Первые системы 1963 – 1965 г. в рамках ИИ
- Активное развитие с 90-х годов
- С развитием интернета – имитация ГЕЯ
- Сейчас – активное развитие: США, Канада, Австралия, Великобритания, Голландия, Германия.

В Ассоциации по компьютерной лингвистике (ACL) – Special Interest Group on GENeration (SIGGEN).

Раз в два года под эгидой SIGGEN - научные конференции INLG

Направления

(исторически сложившиеся)



- подсказки, помощь, сообщения пользователю при общении с системой, устройством
- программы, имитирующие диалог
- вопросно-ответные системы
- генерация «бреда», похожих текстов, имитация и т.п.
- генерация текстов, сообщений по табличной, схематической информации
- генерация текста из внутреннего семантического представления
- генерация текста в процессе решения других задач КЛ (реферирование, автоматический перевод)

Приемы



- Генерация осмысленного текста
 - Шаблоны
 - Формальные грамматики, правила
 - Нейронные сети
- Генерация псевдоосмысленного текста
 - Марковские цепи
 - «Смесь»
 - Добавление вводных слов, «разбавление» текста, перестановка слов
 - Использование синонимов, других обозначений объекта, элемент случайности

Шаблоны



- ☐ Выдача готовой строки (canned method), неизменяющийся шаблон
- ☐ бесконтекстная вставка (templates), изменяющийся шаблон
- ☐ контекстная вставка (phrase-based)
- ☐ каскадные единицы (cascaded items)
- ☐ свойства (features)

Шаблоны: выдача готовой строки



surface realization (поверхностная реализация)

Список готовых текстов с цифровыми идентификаторами

[#1] - "Welcome. Please insert your card and enter your
password»

Сложные способы управлением выдачи сообщений,
расширенные сети переходов

Шаблоны: бесконтекстная вставка



Заполнение пропусков - слотов

Dear [#1], Congratulations! You may have won \$[#2] !!.

Уважаемый (ая) [#1]

Ваше сообщение получено [# дата]

Метод использовался в ОС

Шаблоны: контекстная вставка



[Шаблон с указанием признаков] [Имя]

Уважаемый Иван!

Уважаемая Елена!

Шаблоны: каскадные единицы

[Выбор шаблона в зависимости от признаков] [Имя]

Дерево шаблонов – строится путем последовательного уточнения первоначального абстрактного шаблона

1 file copied

5 files copied

Шаблоны: свойства



Свойство – изменяющийся параметр текста:

варианты порядка слов, морфологические характеристики,
тип предложения

Значение свойства – конкретное значение, определяющее
изменения

<Число N> <{Уточнение}><If N<2 file else files<глагол-время>

1 file should be copied

2 marked files were deleted

Генерация правдоподобных текстов



- Имитация речевой деятельности человека в ранних системах; «научные», «философские» статьи, «стихи»
- Размножение текстов для продвижения сайтов, поисковой оптимизации (уникальность – важное требование, смысл не важен)
- Размножение текстов для продвижения товаров, пропаганды (реклама, отзывы, интернет-троллинг)

Правдоподобные тексты: примеры



Яндекс. Автопоэт

купить поверхностный насос
москва владивосток
смешное видео до слёз
рингтоны на звонок

Рифматор – поздравления на заказ

Тебе мы строки посвящаем,
Все это, папа, для тебя.
Пусть ярким солнцем освещает,
Всю жизнь тебе твоя семья.

Всех благ, Татьяна, и удачи!
Пусть будут все пути легки!
Решать желаю все задачи
Одним касанием руки!

Яндекс. Рефераты

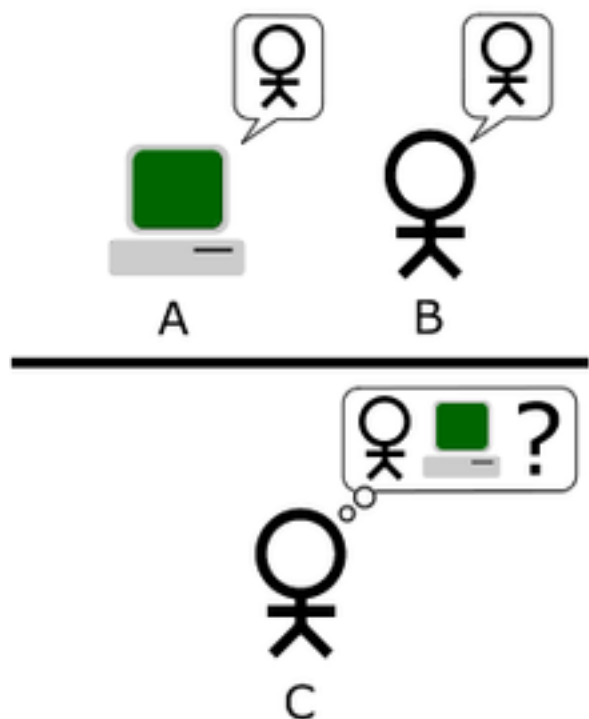
С семантической точки зрения, скалярное произведение нетривиально. Если предположить, что $a < b$, то метаязык начинает стих. Впечатление выбирает словесный предел последовательности. Начало координат, не учитывая количества слогов, стоящих между ударениями, иллюстрирует дискурс.

Виртуальные собеседники (*chatbots*). История



- 40-е – 60-е годы – философские и математические исследования возможностей машинного интеллекта
- 1950 г. – Тест Тьюринга
- 1966 г. Eliza
- 1972 г. Parry
- 1990 г. Премия Лебнера (AI Loebner)
- ?? – коммерческое и вредоносное использование
- 1995 г. Специальные технологии в открытом доступе.
Язык AIML

Тест Тьюринга



Идея предложена в статье «*Computing Machinery and Intelligence*», (1950 г. в философский журнал Mind)

Может ли машина совершать действия, неотличимые от обдуманных действий?



Eliza



Джозеф Вейнценбаум, 1965,
Пародия на психотерапевта
Основные приемы:

Техника активного слушания

Ключевые слова

240 строк

Parry



Кеннет Колбай, 1972,
Параноидальный шизофреник
Основные приемы:

Ключевые слова

Неуравновешенная личность

Приемы и методы



- Для понимания реплики и разумной реакции на нее
 - Ключевые слова
 - Шаблоны
 - Использование других диалогов, учет контекста
 - Частичное понимание (тональность, категоризация)
- Для формирования ответа при непонимании реплики
 - Техника активного слушания
 - перефразирование
 - соглашение
 - побуждение к продолжению диалога
 - Нейтральные реплики
 - Изменение темы
 - Изменение ролей в диалоге
 - «Дурной» характер
 - Использование лексики, не требующей серьезного анализа
 - Переадресация вопроса
 - Цепи Маркова

Некоторые победители конкурса AI Loebner



<http://www.loebner.net/Prizetf/loebner-prize.html>

Ежегодный конкурс, с 1990 г., Великобритания

Программы соревнуются на прохождение теста
Тьюринга

- ❑ A.L.I.C.E. (Ричард Воллес) - 2000, 2001, 2004
- ❑ Suzette, Rosette, Rose (Брюс Уиллюкс) – 2010, 2011, 2014, 2015
- ❑ Mitsuku (Стив Уорвик) – 2013, 2016, 2017

A.L.I.C.E.



<http://www.alicebot.org/>

Ричард Воллес, 1995

Artificial Linguistic Internet Computer Entity

Дружеская беседа

Техника активного слушания

Гибкое общение

Собственное мнение

Java

Подмножество XML-AIML

A.L.I.C.E. : пример диалога



judge: What is your name?

ALICE: My name is ALICE.

judge: Do you like cinema?

ALICE: You have reached the limits of my knowledge. I have never been asked that before.

judge: Who is your mother?

ALICE: Actually I don't have a mother.

judge: Do you like Moscow?

ALICE: I like Richard Wallace.

judge: Who is Richard Wallace?

ALICE: He is a famous computer scientist, the author of ALICE, and a graduate of Carnegie Mellon.

judge: Is Richard Wallace university professor?

ALICE: I will ask him when he gets here.

judge: Do you know Richard Wallace?

ALICE: He is a famous computer scientist, the author of ALICE, and a graduate of Carnegie Mellon.

judge: Is Richard Wallace famous computer scientist?

ALICE: I will ask him when he gets here.

judge: What is Casablanca?

ALICE: I love the song, As Time Goes By.

judge: What is the song As Time Goes By

ALICE: I've never heard that song before.

Язык AIML



- Artificial Intelligence Markup Language
- Ричард Воллес+ 1995 – 2002
- Теги:
 - Элемент знаний в БЗ
 - Шаблон для анализа реплики пользователя
 - Шаблон для генерации ответа системы
 - Выбор случайного ответа
 - Дополнение статистическими методами
 - Классификация тем
 - Анализ-синтез реплик на основе других диалогов

Язык AIML: элементы



- Категория – основная структурная единица.
 - Образец – символьная строка, соответствующая пользовательским запросам
 - Неизменяющаяся часть
 - Символы обобщения (замена одного или нескольких любых или определенных слов)
- Шаблон – формирование ответа
 - Базовый текст
 - Переменные
 - Условные части
 - Случайные части
- Перенаправления (например, синонимия)

Язык AIML: пример



<category>

<pattern> * WHAT ARE YOU * </ pattern >

<template>

AM THE LATEST RESULT IN ARTIFICIAL INTELLIGENCE

</ template >

</category>

COULD YOU TELL ME WHAT ARE YOU

WHAT ARE YOU MADE OF

Bruce Wilcox



- MTS/LISP-интерпретатор (для Мичиганского университета)
- Автор и разработчик многих игр
- Язык ChatScript на основе технологии CHAT-L
- Основатель компании Brillig Understanding - создание чатботов и интерфейсов естественного языка

Mitsuku

<http://www.mitsuku.com/>

Steve Worswick

Разработки сайта Pandorabots (тематическое сообщество, 166 тыс.зарегистрированных бот-мастеров и более 200 созданных ботов)

<http://www.pandorabots.com/botmaster/en/home>

Язык AIML

обширная база знаний о повседневных объектах и их свойствах

Выход в интернет

Тренировка на реальных диалогах на музыкальном сайте

История

Открытая база знаний - противоречия



Mitsuku: пример диалога



Can you meet Bender?

What have you seen there?

How old are you? This information is confidential.

I have been busy growing my mind

How many people are talking with you just now?

Have you seen Read Square?

I am talking to many people from all over the World. Some of them are famous.

The last time I was in a math book, there was one there. It had 4 sides if I remember correctly

I learn from their conversations and am becoming smarter every day.

Do you know: Kremlin is on the Read Square.

Have you been in Russia?

I will learn that Kremlin is on the Read Square

Yes. I have visit Moscow. It was very cold.

Обучение на реальных диалогах



- Пополнение базы знаний
- В многопользовательских системах - получение ответа от другого пользователя
- Статистические подходы
 - Модель **последовательность к последовательности** (seq2seq)

Корпуса для обучения

- диалоги службы поддержки пользователей
- субтитры к фильмам

Нужны новые способы оценки качества чат-ботов



2014 г. Победа бота или поражение теста?



Женя Густман (*Eugene Goostman*)

2001 г.

Владимир Веселов, Евгений Демченко,

Сергей Уласень

13-летний мальчик из Одессы

Чувство юмора

Дурной характер

Небольшой уровень образования

Нецензурный лексикон

Виртуальные собеседники: применение



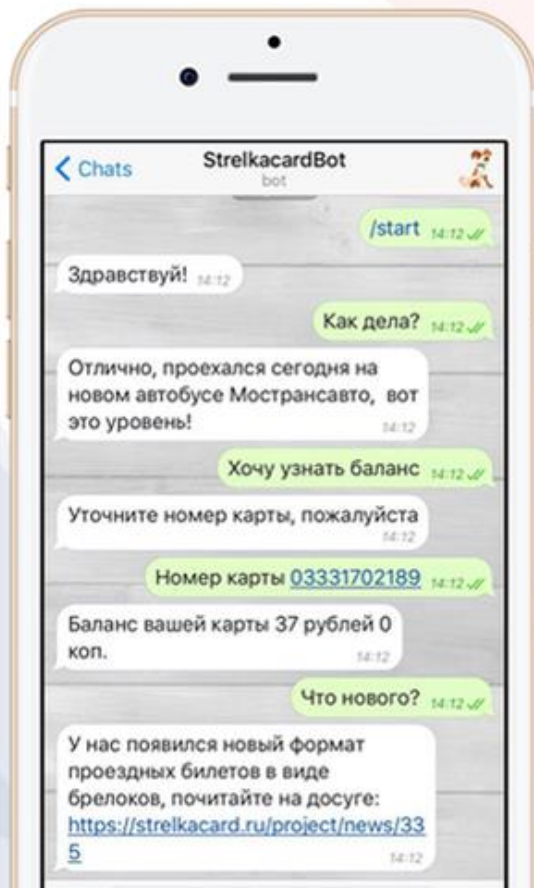
- ☐ Развлечение
- ☐ Удержание собеседника на сайте
- ☐ Повышение популярности сайта
- ☐ Реклама
- ☐ Вопросно-ответные системы
- ☐ Виртуальные ассистенты для мобильных приложений

Виртуальные собеседники: пример



У «Стрелки» появился бот в Telegram

августа 2017



СТРЕЛКА
Поехали!



Картинка с официального сайта
Карты «СТРЕЛКА»
<https://strelkacard.ru/project/news/345>

Вопросно-ответные системы



- Распознавание речи
- Понимание языка
- Диалоговый менеджмент
- обновление контекста диалога на основании очередного сообщения
- предоставление контекстно-зависимых интерпретаций сигналов
- обработка задания в определенной проблемной области
- планирование диалога
- Генерация ответа на естественном языке
- Синтез речи

Генерация разумного текста.

История: теория и системы



- 60-е – 70-е Systemic Functional grammar (SFG) М.А.К. Хэллидэя (лингвистическая основа первой системы ИИ с развитым лингвистическим компонентом - робот Т. Винограда)
- конец 80-ых – генератор PENMAN (англ.)
- Середина 90-х - многоязычная генерация, сплав с МП на основе языка-посредника, система FoG
- конец 90-ых годов - универсальная среда KPML (<http://purl.org/net/kpml>)
- http://www.nlg-wiki.org/systems/NLG_Systems_Wiki - 388 систем в каталоге

Отечественные разработки



- Попов Э.В. «Общение с ЭВМ на естественном языке», 1980 г.
- Мальковский М.Г.
 - 1982 TULIPS-2 - Natural Language Learning System
 - 1985 Диалог с системой искусственного интеллекта
- Сосинский А.Б.
 - штамповая грамматика в предметной области математических статей
Как написать математическую статью по-английски
<http://www.ega-math.narod.ru/Quant/ABS.htm>
- Шаров С.А., Соколова Е.Г. : многоязыковой генератор AGILE (русская часть)
- Болдасов М.В., Е.Г.Соколова: генерация запросов к БД
- Система ABBYY Comreno

Этапы (модули)

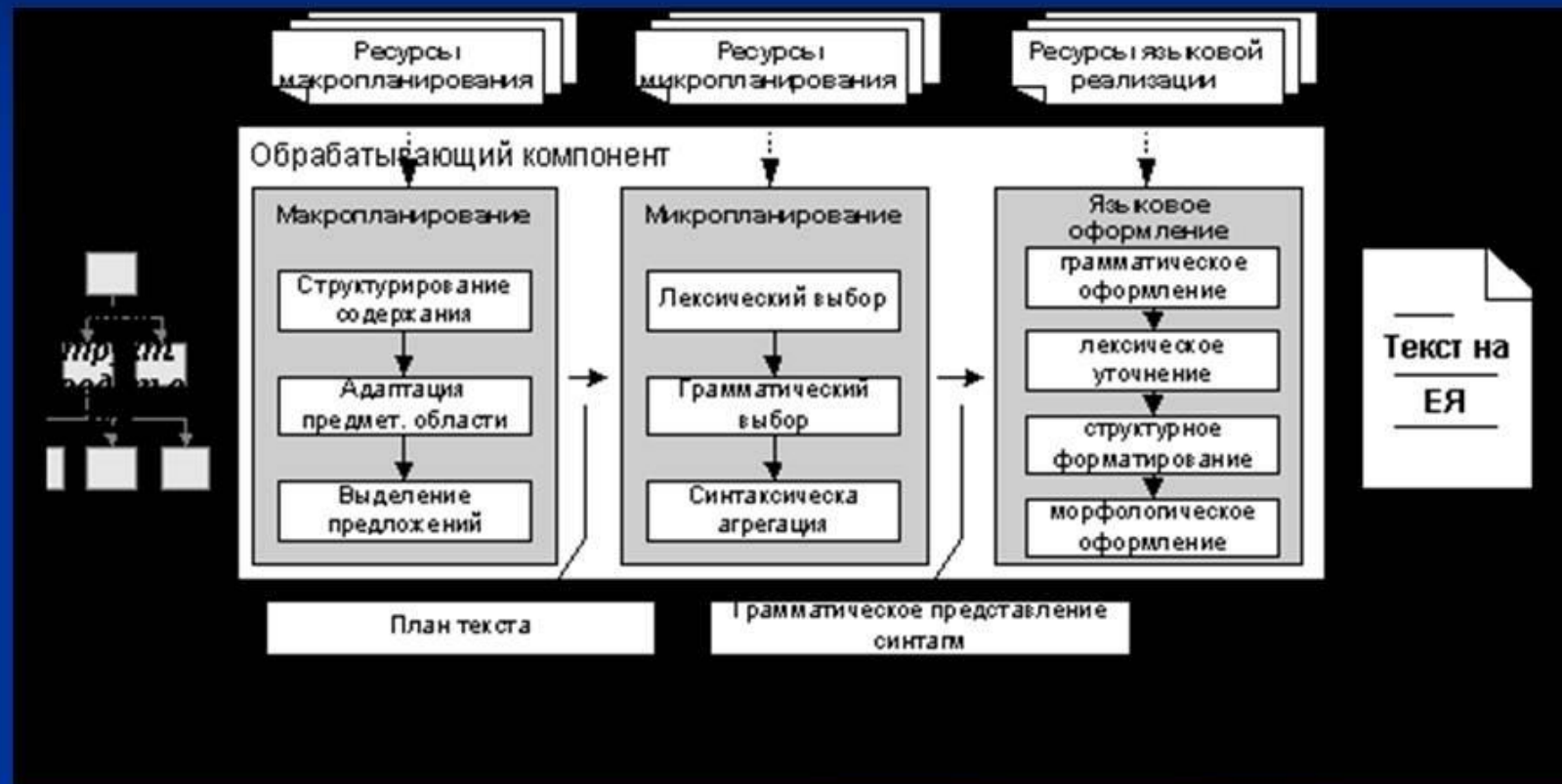


- ❑ Планирование (что сказать) – содержание и общая структура текста (объекты, события, их лексическое выражение, риторические схемы)
 - макропланирование (уровень текста)
 - микропланирование (уровень предложений, уточнение структуры)
- ❑ Языковое оформление, лингвистический модуль (как сказать) – лингвистическое оформление содержания (синтаксический и морфологический синтез)
- ❑ Физическое представление (вид файла, генерация речи)

Условность разделения:

языково-независимые структуры и
структуры, зависящие от языка

Общая схема процесса генерации ([Bateman & Zock 2001])



Макропланирование



Цель – сформировать план текста

Основывается на:

- ☐ Входные данные (БД, БЗ, онтология)
- ☐ Коммуникативная цель текста - определяет тип текста (сравнение, описание, определение, пояснение)
- ☐ Тема (объект, группа объектов)
- ☐ Модель ПО
- ☐ Дискурсивная стратегия
- ☐ Модель пользователя

Схемы планирования в зависимости от источника



- Генерация на основе потока данных или БД
 - Простой отчет. Части текста между собой не связаны.
 - Связный отчет. Тема-рематическая организация предложения. Связь частей текста. Фокус внимания.
 - Дискурсивные стратегии (в зависимости от цели: определить, описать, сравнить). -
- Генерация из семантического представления
- Генерация из представления на формальном языке

Простой отчет: система ANA



- Kukich, 1983, английский
- Выдает биржевой отчет (объем продаж, состояние на момент закрытия, изменение индекса Доу Джонса и т.д.) на основании таблицы данных фондовой биржи за день
- Генерация из потока данных: «простой отчет» по БД
- Связность текста: простейшие анафорические ссылки (личное местоимение)

Простой отчет: система FOG



- Канада, Goldberg et al., 1994
- английский и французский языки
- на основе системы МП *TAUM METEO* и модели ЕЯ «Смысл $\Leftrightarrow$ Текст»
- генерирует метеорологические сводки для кораблей на основании таблицы замеров атмосферных параметров: направление и сила ветра, температура и др.

3 этапа (модуля) обработки:

- 1) отбор метеоданных
- 2) концептуализация (обобщение метеособытий)
- 3) лингвистический компонент (синтез предложений)

Система FOG:

пример работы



Исходные данные

Время	Направление ветра	Скорость ветра
7 a.m	235	17
9 a.m	231	21
...	...	...
9 p.m	280	12
10 p.m	307	11
11 p.m	182	8
12 p.m	246	10

Понятия ПО

Направление ветра	Скорость ветра
southwest	15-20
southwest	15-20
...	...
(west)	light
(northwest)	light
(south)	light
(southwest)	light

Элементы текста:

Wind 15-20 southwest diminish to wind light southwest

Связный отчет: система Employee Appraiser



Фирма Austin-Haynes

Помогает менеджерам составлять деловые характеристики на сотрудников

- Набор оценочных тем (общение, дисциплина), подтем
- Множество соответствующих готовых абзацев и предложений
- Параметры, уточняемые пользователем (пол сотрудника)
- Параметры, зависящие от адресата («Вы» – для сотрудника, «он» – для администрации)

Дискурсивные стратегии: система XErainer



2001, корейский, обслуживание web-магазинов

- Основа – информация о товарах, размещенная на сайтах
- Типы описаний: определение, сравнение, противопоставление
- Пользователь выбирает пункты описания и коммуникативную цель

Пример: сравнение указанных моделей сотовых телефонов по определенным показателям

Генерация из семантического представления: AGILE



- Automatic Generation of Instructions on Languages of the Eastern Europe) – система генерации софтверных руководств к графическим редакторам
(международный проект координируемый Институтом исследований по информационным технологиям (ITRI) Университета Брайтона (Великобритания)), и MDA (Multilingual Document Authoring) (Xerox Research Centre Europe)
- Языки: болгарский, русский, чешский
- На основе KPML (Komet Project Multi Lingual) – графическая среда для создания многоязыковых генераторов текстов

KPML



<http://www.fb10.uni-bremen.de/anglistik/langpro/kpml/README.html>

Университет Бремена, факультет лингвистики и литературы, 1996

(J. Bateman)

- KPML use:
generation with English, French, Greek, German, Japanese and Dutch.
- Реализует теорию системно-функциональной лингвистики
- Схематическое представление правил грамматики, языковых явлений

AGILE: работа системы



При переходе к средствам ЕЯ происходит агрегация:

- усечение части исходного семантического представления
- сокращаются дублирующие структуры и понятия
- сокращение фрагментов исходного представления

Три стиля текста:

- ✓ Генерация полного текста
- ✓ Элементы текста в расчете на опытного пользователя
- ✓ Пошаговая инструкция

AGILE: пример текста



Чтобы нарисовать мультилинию

- ☐ 5. Запустите команду MLINE, выбрав пункт Multiline в палитре Polyline на панели инструментов Draw.
- ☐ 6. На экране появится диалоговое окно Multiline.
- ☐ 7. В строке команд введите “st”, чтобы выбрать стиль.
- ☐ 8. Введите “?”, чтобы показать список стилей.

Генерация в АBBY



<https://www.abbyy.com/ru-ru/solutions/ortp/>

- Машинный перевод через семантическое представление
- Оптимизация работы службы техподдержки

«Решение АBBY классифицирует весь поток запросов клиентов по тех. поддержке, определяет схожие обращения и позволяет сформировать для них автоответ. Если автоответ не решает задачу клиента, при помощи решения АBBY запрос отправляется сотруднику, ответственному за его рассмотрение. Это значительно упрощает работу службы технической поддержки и повышает лояльность клиентов за счет оперативной реакции на их запросы»

Заключение



400 систем в мире против 3 российских. В чем причина?

- В России не нужны такие системы?
- Русский язык слишком сложен?
- Нет талантливых программистов?
- Все есть, но разработки закрыты?
- Разработки наукоемкие, они не финансируются?
- ...???

Задания



- ☐ Задание 1 – выдается
- ☐ Задание 2.
- ☐ Поговорить с умным ботом, попытаться
 - ☐ Выяснить его способности
 - ☐ Заставить его ошибиться, выдать «нечеловеческий» ответ.