Автоматизированная обработка естественного языка

Natural Language Processing

NLP: истоки

- **возникло в конце 60-х гг.**
- развивалось в рамках дисциплины «искусственный интеллект».
- АОЕЯ — разработка методов, технологий и конкретных систем, обеспечивающих общение человека с ЭВМ на естественном или ограниченном естественном языке.

Проблема организации взаимодействия с компьютерными

- Решение этой проблемы коммуникации шло по двум основным путям.

1 - адаптация языков программирования и операционных систем к конечному пользователю.

2 - разработка систем взаимодействия с ЭВМ на естественном языке или каком-то его ограниченном варианте.

NLP и ЛИНГВИСТИКА

- Фонология (звуки речи)
- Морфология (структура и форма слов ЕЯ)
- Синтаксис (структура и функции предложений)
- Семантика (смысл языковых высказываний)
- Прагматика (значение высказываний)
- Социолингвистика Психолингвистика
- Лексикография (описание лексикона ЕЯ)
- Прикладная лингвистика

NLP: МАТЕМАТИКА и ИНФОРМАТИКА

- Математическая лингвистика
- Квантитативная лингвистика (изучение языка/речи количественными методами)
- Теория формальных языков и грамматик – возникла из порождающих грамматик Н.Хомского (50-е гг.), для анализа синтаксических структур ЕЯ
- Теория алгоритмов
- Информатика (Computer Science)

NLP и ИСКУССТВЕННЫЙ ИНТЕЛЕКТ

- Междисциплинарный характер области ИИ: составная часть Computer Science , пересечение (по задачам и методам) с АОТ
- Задача ИИ – компьютерное моделирование интеллектуальных функций

- Первая известная программа ИИ по обработке ЕЯ – Система Т. Винограда (70-е годы);
- Пример диалога : Pick up a big red block. (человек) OK (машина) Is there a large block behind a pyramid? Yes, Three of them. Grasp the pyramid. I don’t understand, which pyramid you mean

ОСОБЕННОСТИ ЕЯ

- ЕЯ – сложная система знаков, возникшая для обмена информацией в процессе человеческой деятельности и постоянно изменяющаяся вместе с ней
- Две стороны знака: означаемое – означающее
- Сложности ЕЯ
  - комбинаторная система яз. знаков
  - многоуровневость системы ЕЯ
  - каждый уровень (подсистема) – правила сочетания знаков
  - взаимосвязь уровней
- Разнообразие языков и языковые универсалии

ОСОБЕННОСТИ ЕЯ: УРОВНИ

1. **Фонологический**: звуки (фонемы)/ буквы – незначащие единицы , средство различения др. единиц

2. **Морфологический** – слова (словоформы)
- подуровень морфем

3. **Синтаксический** – предложения (фразы) ЕЯ
- подуровень словосочетаний
- надуровень сверхфразовых единств (≈ абзацев) – предложений, объединяющихся по смыслу

⇒ возможность построить практически бесконечное число высказываний (смыслов)

ДОПОЛНИТЕЛЬНЫЕ УРОВНИ ЕЯ:

- **Семантический** : набор элементарных единиц – сем
- **Лексический** : множество лексем (лексикон)
- **Дискурсивный** (уровень связного текста): схематические структуры текстов (патентные формулы, деловые письма и т.п.)

- Сложность системы ЕЯ
- Взаимосвязь всех уровней
- Нестандартная сочетаемость (синтактика) единиц ЕЯ на всех уровнях
- Большая системность (число уровней)
- Асиметрия связи единиц и выражаемых ими смыслов: полисемия, синонимия, омонимия

⇒ невозможность единожды создать лингв. процессор

Сложность ЕЯ ⇒МОДУЛЬНОСТЬ ЛИНГВ. ПРОЦЕСОРОВ

- Графематический анализ
- Морфологический анализ
- Постморфологический анализ: разрешение морфологической омонимии
- Предсинтаксис: сегментация текста на предложения
- Синтаксический анализ предложений
- Семантический и прагматический анализ

Архитектура систем NLP

- блок **анализа** речевого сообщения пользователя,
- блок **интерпретации** сообщения,
- блок **порождения** смысла ответа,
- блок **синтеза** поверхностной структуры высказывания,
- **диалоговый** компонент

Блок анализа

- **морфологический** анализ словоформ
- **синтаксический** и **семантический** анализ предложений.

Блок порождения смысла

- **определение информации**, которую следует передать пользователю,
- предполагаемое **членение** информации на «**порции**», соответствующие предложению;
- определение **последовательности** «порций» смысла;
- построение **семантического представления** отдельных предложений

Блок синтеза поверхностной структуры высказывания

- **упаковка** семантического представления высказывания в **синтаксические структуры предложения**.
- Здесь играют существенную роль категории коммуникативной организации смысла высказывания — тема, рема, данное, новое

Классические задачи обработки текстов

- Информационный поиск (IR)
- Извлечение информации (IE)
- Вопросно-ответные системы (QA)
- Классификация и кластеризация
- Автоматическое аннотирование и реферирование
- Диалоговые системы
- Машинный перевод

Современные речевые технологии

1. Распознавание речи
2. Синтез речи по тексту
3. Выделение ключевых слов в слитной речи

– **Siri**: интеллектуальный ассистент на iPhone

ПРИКЛАДНЫЕ ЗАДАЧИ NLP

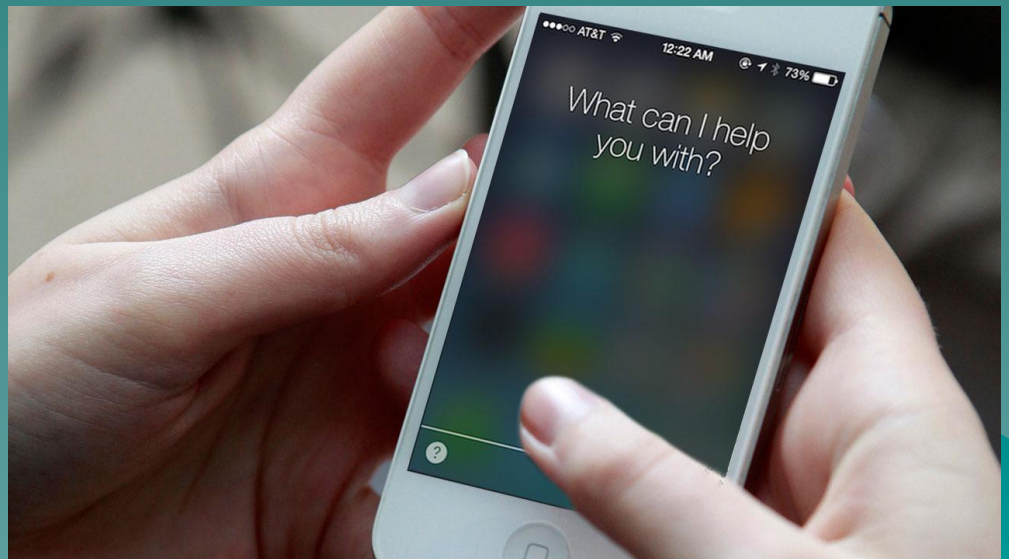
4. Определение языка сообщений
5. Идентификация диктора
6. Определение эмоционального и физического состояния человека по его голосу.
7. Шумоочистка
8. Разделение дикторов
9. Music Spotting

Диалоговые системы

Пользователь: Мне нужен самолет из СПб до Бостона, прибытие в 10 утра.
Компьютер: На какое число?
Пользователь: На завтра.
Компьютер: Доступные рейсы на завтра...
<список рейсов>

Siri

- англ. Speech Interpretation and Recognition Interface
- персональный помощник и вопросно-ответная система, адаптированная для iPhone OS.
- использует обработку естественной речи, чтобы отвечать на вопросы и давать рекомендации
- приспосабливается к каждому пользователю индивидуально, изучая его предпочтения в течение долгого времени.



Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

Dracula

What is Watson?

IBM WATSON

- Open Domain Question-Answering Machine
- Given
  - Rich **Natural Language Questions**
  - Over a **Broad Domain of Knowledge**
- Delivers
  - **Precise Answers:** Determine what is being asked & give precise response
  - **Accurate Confidences:** Determine likelihood answer is correct
  - **Consumable Justifications:** Explain why the answer is right
  - **Fast Response Time:** Precision & Confidence in <3 seconds
  - At the level of human experts
- Proved its mettle in a televised match
  - Won a 2-game Jeopardy match against the all-time winners
  - viewed by over 50,000,000

Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jurafsky

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum m
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Create new Calendar entry

Извлечение информации

Сдаю в г.Пушкин комнату 11 м2 в 3-х комнатной квартире без соседей, в квартире никто не живет, коридор 15 м2, кухня 15 м2, балкон,стиральная машинка,холодильник,телевизор.До метро Купчино 15 минут на маршрутке и автобусе, остановка рядом с домом,до станции 20 минут.Аренда 9000 руб+2000 залог+свет+вода.Тел.8-921-898-59-60 Юлия

↓

Город	Пушкин
Количество комнат	3
Метро	Купчино
Цена	9000 руб.

Извлечение информации

- Цель: представить коллекцию текстовых документов в структурированном виде
- Мотивация:
  - Запросы сложного характера: Квартира рядом с метро Петроградская ценой не больше 30 тыс. руб.
  - Статистические запросы: Как изменились цены на квартиры в Купчино за последний год?

Information Extraction & Sentiment Analy

Information Extraction & Sentiment Analysis

Attributes:

- zoom
- affordability
- size and weight
- flash
- ease of use

Information Extraction & Sentiment Analy

Attributes:
zoom
affordability
size and weight
flash
ease of use

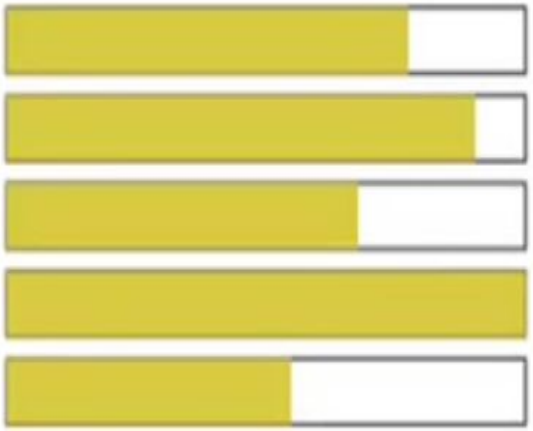
Size and weight

- nice and compact to carry!
- since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- the camera feels flimsy, is plastic and very light in weight have to be very delicate in the handling of this camera

Information Extraction & Sentiment Analysis

Attributes:

- zoom
- affordability
- size and weight
- flash
- ease of use



Size and weight

✓ • nice and compact to carry!

✓ • since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!

✗ • the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Пример

I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long.

Что мы здесь наблюдаем?

Пример

I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long.

Что мы здесь наблюдаем?

- Целевой объект
- Аспекты объекта
- Мнения (положительные и не очень)

Machine Translation

- Fully automatic

Enter Source Text:

这 不过 是 一 个 时间 的 问题 .

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translator

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

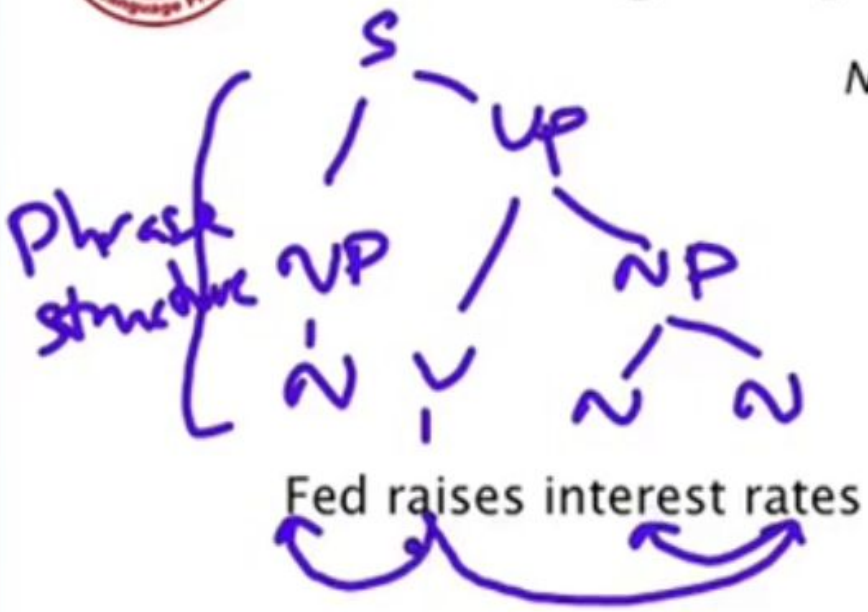
Почему анализировать тексты трудно?

Хочу обнимать тебя часами

- Мне улыбнулась хозяйка фермы, которая красовалась возле магазина.
- Squad helps dog bite victim.

Ambiguity is pervasive

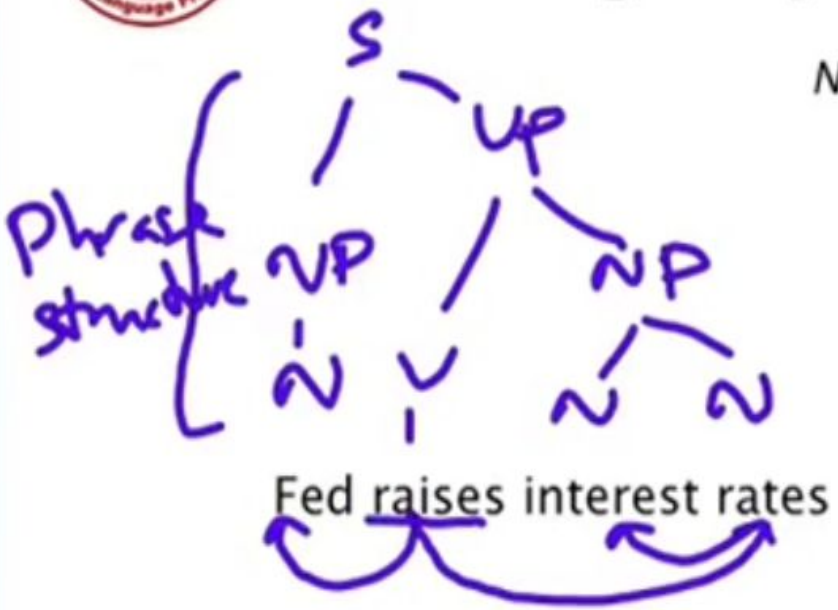
New York Times headline (17 May 2000)



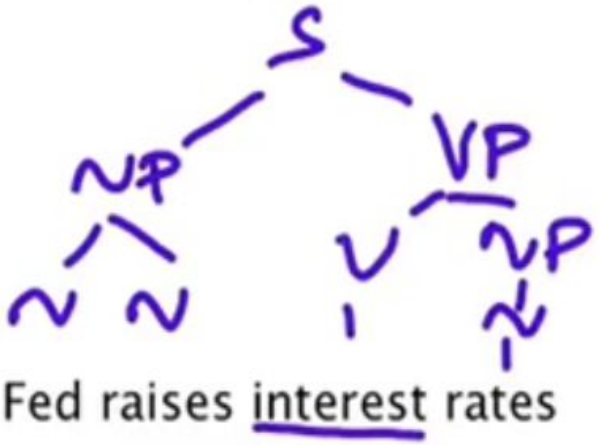
Fed raises interest rates

Fed raises interest rates 0.5%

Ambiguity is pervasive



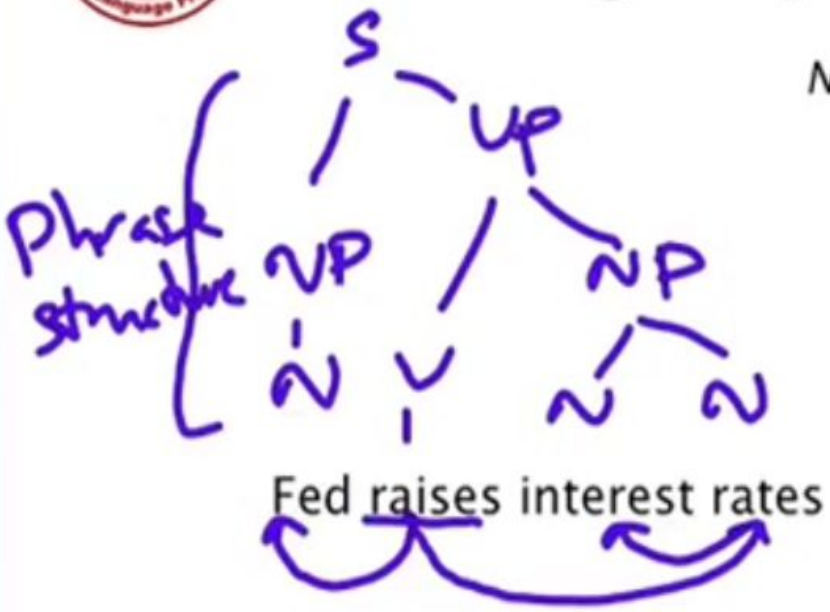
New York Times headline (17 May 2000)

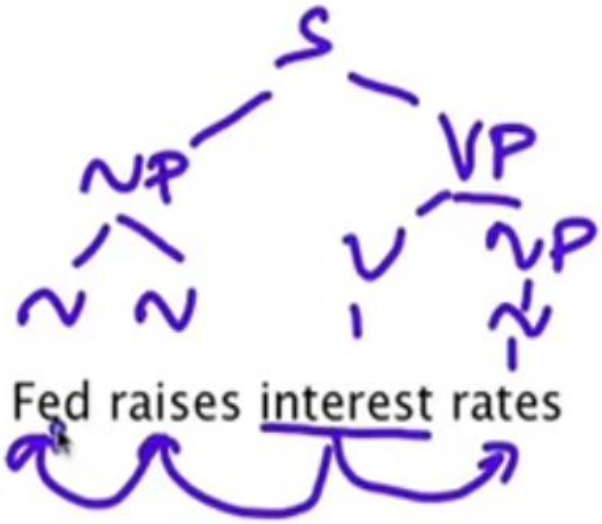


Fed raises interest rates 0.5%

Ambiguity is pervasive

New York Times headline (17 May 2000)





Многозначность

- Морфологическая
  - часть речи
  - мой (-- нос, -- руки)
  - look (look at me, have a look)
- Синтаксическая
  - мужу изменять нельзя
  - мать любит дочь
  - Flying planes can be dangerous

Алгоритмы определения частей речи (part of speech tagging)

Синтаксический разбор (parsing)

Многозначность

- Лексическая (семантическая)
  - Омонимия (ключ)
  - полисемия (платформа)
  - семантическая многозначность (лиса)
- Прагматическая
  - тот человек в толпе...
  - You have a green light

разрешение лексической многозначности (word sense disambiguation)

Dan Jurafsky

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.
Mary and Sue are mothers.

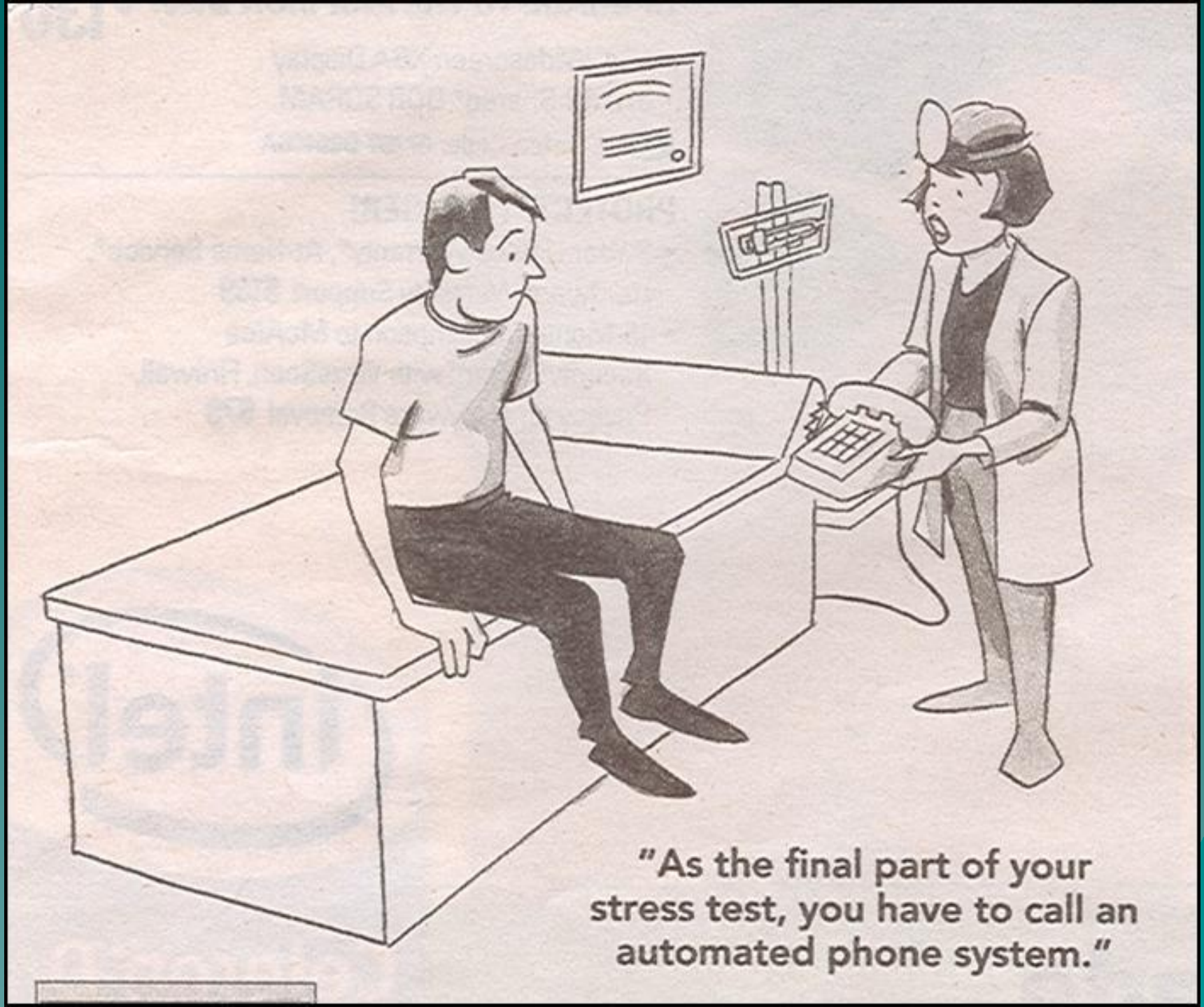
tricky entity names

Where is *A Bug's Life* playing ...
Let It Be was recorded ...
... a mutation on the *for* gene ...

But that's what makes it fun!

Понимание

- Тест Тьюринга (Премия Лёбнера)
- Китайская комната
- ELIZA (1966)
  - „I am unhappy."
  - „How long have you been unhappy?"
  - „Six months. Can you help me?"
  - „What makes you think I help you?"
  - „You remind me of my sister."
  - „Can you tell me more about your sister?"
  - „I like teasing my sister."
  - „Does anyone else in your family like teasing your sister?"
  - „No. Only me."
  - „Please go on."


"As the final part of your stress test, you have to call an automated phone system."