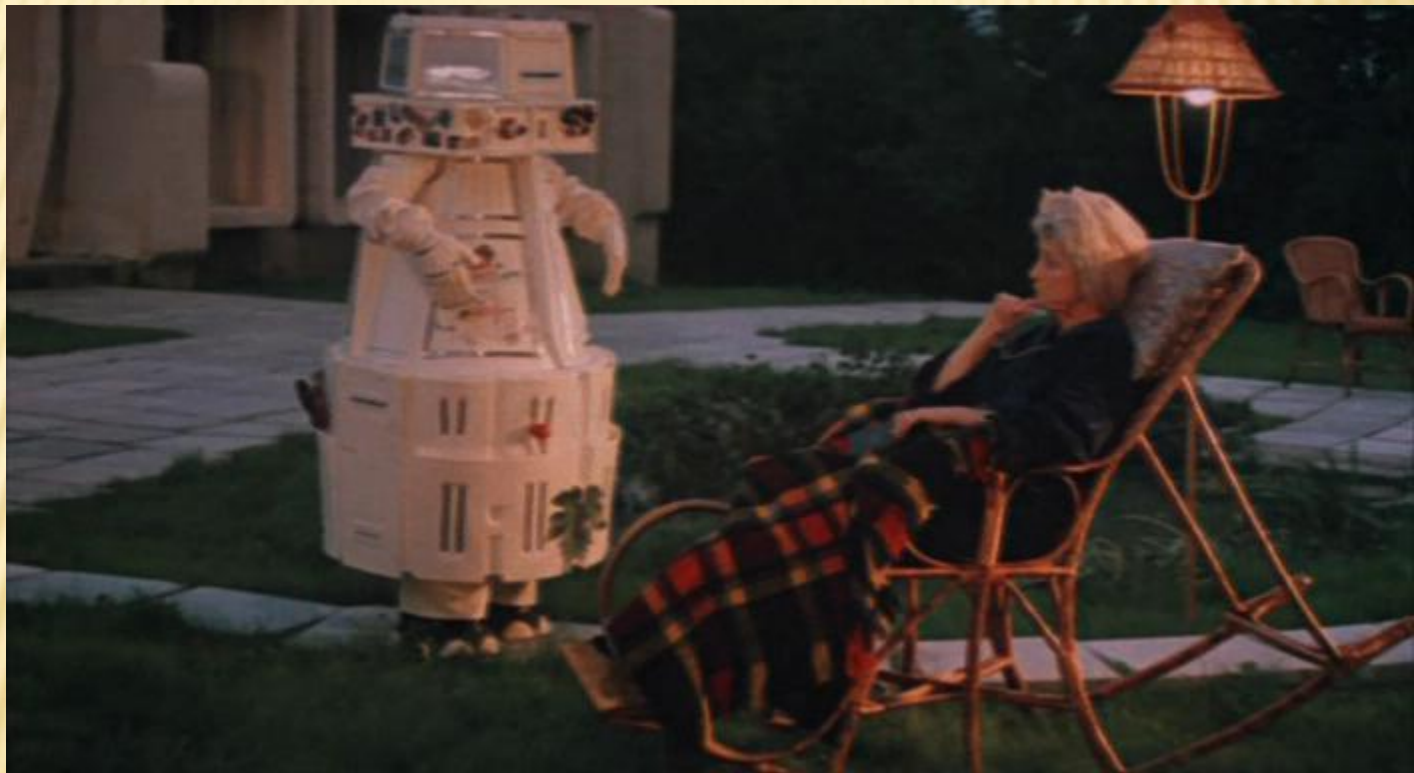


МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА КАК НАУКА БУДУЩЕГО

О.В. Митренина,
доцент кафедры математической лингвистики
Санкт-Петербургского университета
mitrenina@gmail.com

РОБОТЫ XXI ВЕКА, ВИД ИЗ ПРОШЛОГО



НЕКОТОРЫЕ ЧАСТНЫЕ ЗАДАЧИ

1. Machine Translation
 2. Information Retrieval
 3. Information Extraction
 4. Sentiment Analysis
 5. Question Answering
 6. Dialogue
- и пр.

ТЕХНОЛОГИИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

1. Rule-based (правила, составленные экспертами)

- ✗ Слова (род-число-падеж-склонение)
- ✗ Структура предложения (роли слов)
- ✗ Смыслы
- ✗ Звуки

ТЕХНОЛОГИИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

2. Statistical (не требуют знания языка, нужна только математика!)

Машинное обучение (Machine Learning)

Подход был известен в СССР как

"Математическая оптимизация" или

"Теория статистического управления".

МОДЕЛЬ ЯЗЫКА

Распознавание речи:

- ✗ Расцветали яблони и груши
- ✗ Раз цвета ли яблони?

Перевод: «Съешь этот лук»

- ✗ Eat this onion.
- ✗ Eat this bow.

ВЕРОЯТНОСТЬ ПРЕДЛОЖЕНИЯ

Как ее определить?

1) $p(\text{Сегодня выпал снег}) =$

$p(\text{Сегодня}) \times p(\text{выпал}) \times p(\text{снег})$

Манная...

каша / запеканка / крупа

Съешь манную ...

кашу.

ТРИГРАММНАЯ МОДЕЛЬ

$p(\text{Сегодня выпал снег}) =$

$p(\text{Сегодня} \mid * *) \times$

$\times p(\text{выпал} \mid *, \text{Сегодня}) \times$

$\times p(\text{снег} \mid \text{Сегодня, выпал}) \times$

$\times p(\text{STOP} \mid \text{выпал, снег})$

МАЛЕНЬКИЙ СОСЕД В КУПЕ

42 предложения начинались со слова «хочу», в том числе:

- ✗ Хочу собаку — 21 раз
- ✗ Хочу гулять — 12 раз
- ✗ Хочу домой — 6 раз
- ...
- ✗ Хочу спать — 0 раз

ОЦЕНКА МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Maximum-Likelihood Estimates

Нужен корпус.

$p(\text{снег} \mid \text{Сегодня, выпал}) =$

$= c(\text{Сегодня выпал снег}) / c(\text{сегодня выпал})$

$p(w \mid u, v) = c(uvw) / c(uv)$

МОЖЕТ ОБЩАТЬСЯ ШЕСТЬЮ МИЛЛИОНАМИ СПОСОБОВ

