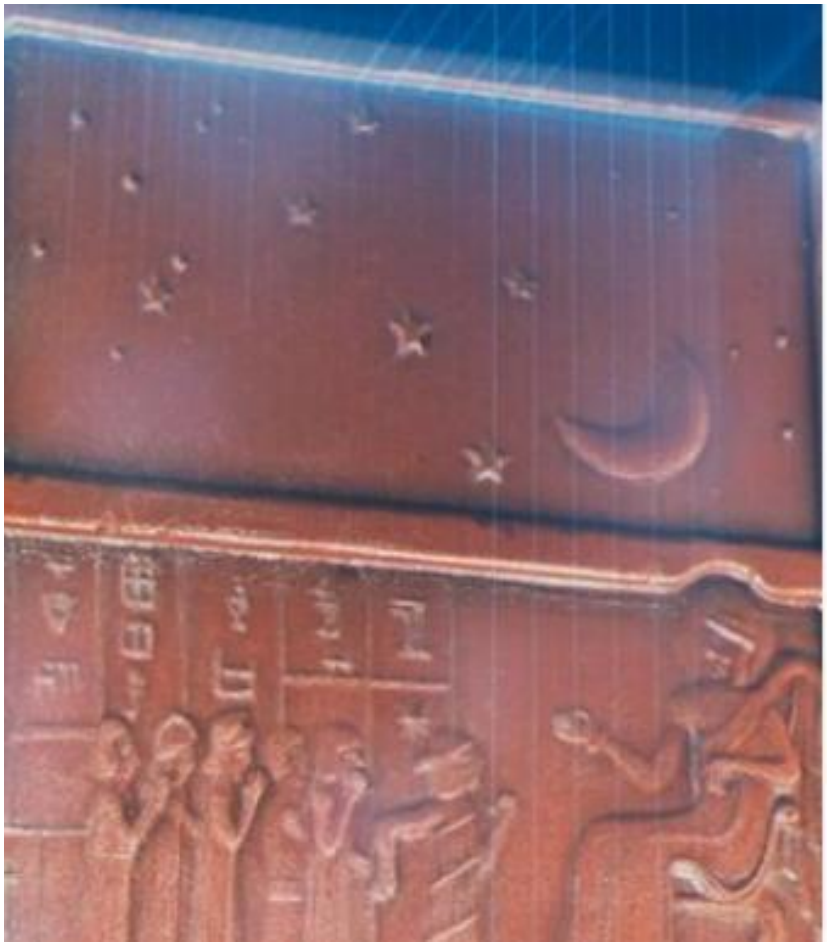
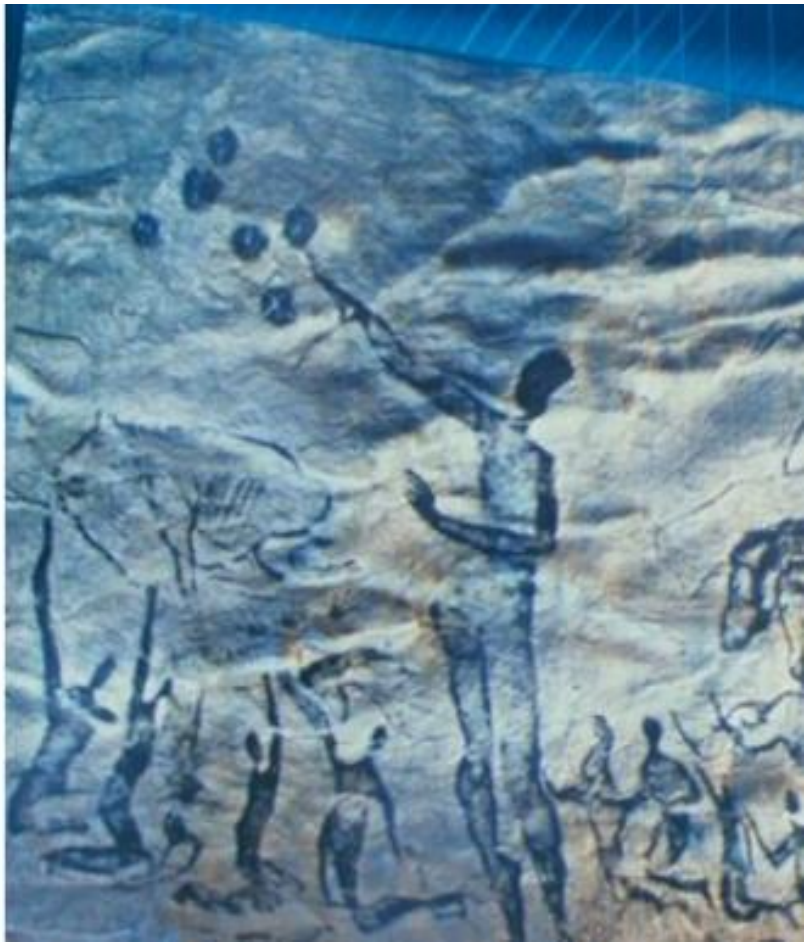




O que vemos aqui?

SCC0633/SCC5908

Processamento de Linguagem Natural

David no Planetário

David e Engenheiro

Metas

2012

- David
  - Tarefas simples
  - Apoio ao humano
  - ...
  - Comunicação “universal”
  - Até decodificação de línguas desconhecidas e perdidas
    - Língua proto-indo europeia, uma das mais antigas que se conhece

Diretor: Ridley Scott

Vários anos atrás...

- **How Revolutionary Tools Cracked a 1700s Code**
  - *Now a team of Swedish and American linguists has applied statistics-based translation techniques to crack one of the most stubborn of codes: the Copiale Cipher, a hand-lettered 105-page manuscript that appears to date from the late 18th century. They described their work at a meeting of the Association for Computational Linguistics in Portland, Ore.*
  - *Kevin Knight, a computer scientist at the Information Sciences Institute at the University of Southern California, collaborated with Beata Megyesi and Christiane Schaefer of Uppsala University in Sweden to decipher the first 16 pages. They turn out to be a detailed description of a ritual from a secret society that apparently had a fascination with eye surgery and ophthalmology.*

http://www.nytimes.com/2011/10/25/science/25code.html

Alguns anos atrás

- *Novos Google Pixel Buds mostram por que a tecnologia é maravilhosa*

  https://www.tecmundo.com.br/produto/122679-novos-google-pixel-buds-mostram-tecnologia-maravilhosa.htm

Recentemente

- GPT (*Generative Pre-training Transformer*)
  - O desenvolvimento das tecnologias de inteligência artificial se tornou cada vez mais sofisticado ao longo dos últimos anos. Computadores com maior poder de processamento vêm adquirindo repertórios abrangentes, dominando diversas áreas do conhecimento e capacidades diferentes. Um exemplo deste desenvolvimento é o GPT-3: sistema desenvolvido pela OpenAI baseado em *machine learning* (aprendizado de máquinas) que possui a capacidade de escrever diversos tipos de gêneros textuais com grande verossimilhança a qualquer trabalho executado por um humano... Da poesia a literatura, do jornalismo ao direito: as possibilidades de criação original por parte do GPT-3 demonstram que seu poder produção é vasto.

    https://olhardigital.com.br/2020/08/25/noticias/gpt-3-o-mais-poderoso-sistema-de-inteligencia-artificial-ja-criado/

No dia a dia

- Corretores ortográficos
- Tradutores automáticos
- Reconhecedores de fala
- Assistentes virtuais
- Sistemas de recomendação
- Etc.

Desafios

- Máquinas que interagem “humanamente”

  Mais do que ser competente, desempenhar!

- Decodificação universal
  - Línguas do passado remoto... e também alienígenas!

- Linguagem...
  - e mente e evolução (e quem veio primeiro)
  - e pensamento e consciência (e a “crise” de David)
  - e comunicação (e o que nos separa dos outros animais)
  - e máquinas (e a “verdadeira” inteligência artificial)
    - e mineração de textos, *big data*, ciência de dados, etc.

Nesta disciplina

- O início desse percurso

SCC5908 Introdução ao Processamento de Língua Natural

&

SCC0633 Processamento de Linguagem Natural

Prof. Thiago A. S. Pardo
Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

Esta disciplina

- Foco desta disciplina
  - **Fundamentação linguística básica**
  - **Representações linguístico-computacionais usuais**
  - **Métodos computacionais essenciais**
  - **Prática**

- Envolve *Computação*, *Linguística* e *Estatística*, pelo menos
  - E também um pouco de *Psicologia*, *Filosofia*, *Física*...

- Objetivo: fornecer ao aluno uma visão geral da área, com fundamento, raciocínio e postura crítica

Ementa

- Visão geral da área de Processamento de Línguas Naturais (PLN) e sua relação com as áreas de Computação e Linguística. História da área e seus principais marcos. Níveis de representação e processamento linguístico: fonética e fonologia, morfologia, sintaxe, semântica, discurso e pragmática. Familiarização e prática com recursos, ferramentas e aplicações de PLN. Abordagens e paradigmas para a resolução de problemas de PLN. Construção e anotação de córpus. Prática de projeto e desenvolvimento de um protótipo computacional de PLN. Apresentação de tópicos relevantes atuais de PLN, assim como modelos e métodos associados.

Pré-requisitos

- Boas noções de programação

- Conteúdo essencial de Inteligência Artificial, com destaque para aprendizado de máquina e representação de conhecimento

  - Se não é o caso, aproveitar a oportunidade para se informar e atualizar sobre esse conteúdo
    - http://aima.cs.berkeley.edu/
    - https://www.cs.waikato.ac.nz/~ml/weka/book.html

Como será a disciplina

- Aulas virtuais síncronas
  - Gravadas e disponibilizadas posteriormente
- Material de apoio
  - Slides
  - Uso de softwares e pacotes especializados
  - Resolução de exercícios
  - Leituras complementares
- Plataforma e-Disciplinas
  - **Quanto mais informal, melhor!**

Avaliação

- **Provas virtuais curtas semanais**
  - Liberadas no e-Disciplinas ao fim de cada aula, podendo ser entregues até o início da aula seguinte
    - Conteúdo da aula e das leituras será cobrado
  - Ocasionalmente podem ser substituídas por exercícios práticos
  - Nota final: média das notas

- Presença verificada indiretamente, pela resolução de exercícios, provas e rendimento individual

Disciplina espelhada

- **Graduação e pós-graduação juntas**
  - Exercício interessante para os dois lados!
    - Atualização e tecnicalidade da graduação
    - Profundidade da pós-graduação

- Horários alinhados à graduação
  - Primeira parte da aula: das 8h10 às 9h50
  - Segunda parte da aula: das 10h10 às 11h50

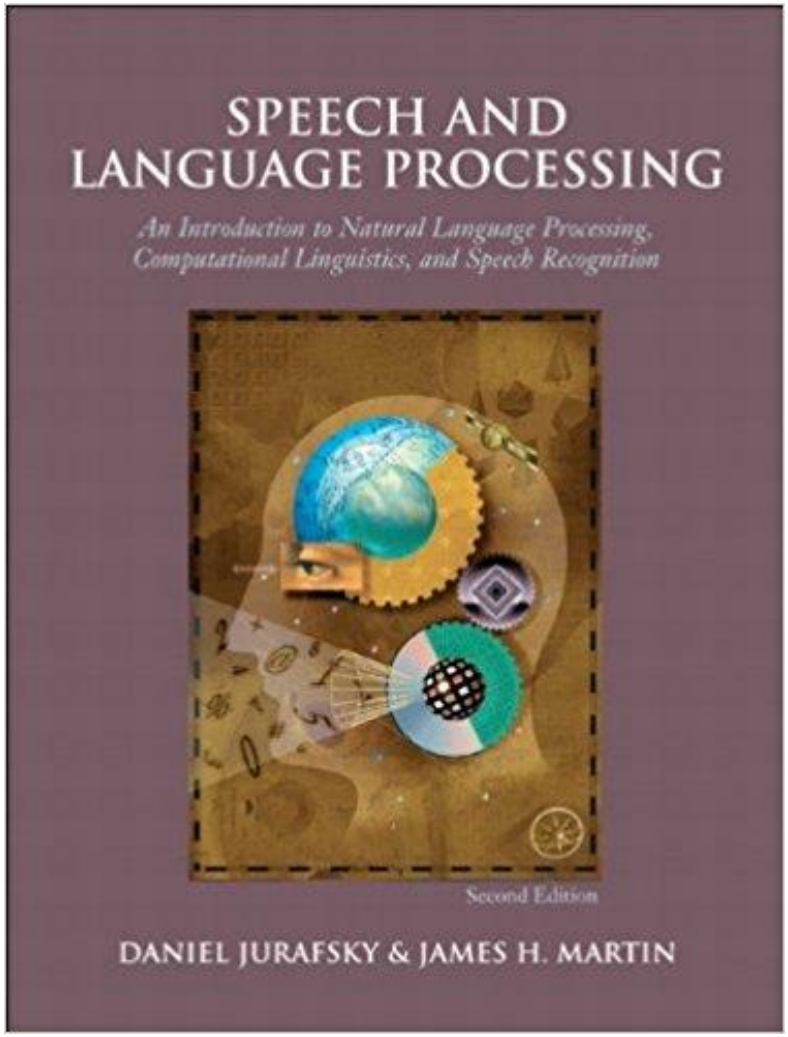
- Somente para graduação
  - Prova REC: 05/agosto

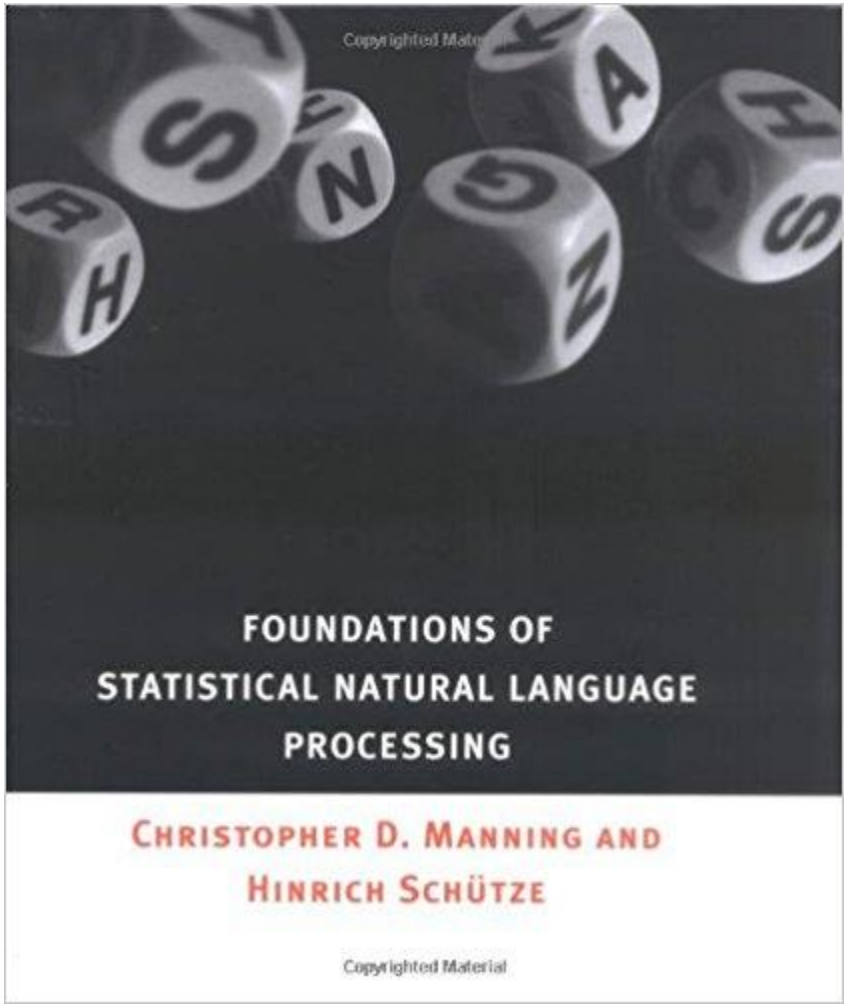
- Somente para pós-graduação
  - Conceito A, se média final >= 8,5
  - Conceito B, se média final >= 7 e < 8,5
  - Conceito C, se média final >= 5 e < 7
  - Reprovado, se média final < 5

Os domínios e a natureza do conhecimento

- Transdisciplinaridade e seus desafios
  - Computação + Linguística, principalmente
    - Mas há muito mais no mundo
  - Temos que vencer o mito do “outro lado”
    - Somos capazes! ☺
  - Desafios
    - Pensamento computacional e algoritmos
    - Modelos linguísticos
    - Estatística
    - Subjetividade
    - Etc.

Bibliografia básica (mas há muito mais)





Muitos capítulos disponíveis online
na página dos autores

Estagiário PAE

- Roney L. S. Santos (doutorando)
  - Horários de atendimento online?
    - Quinta, das 12 às 13h
    - Sexta, das 16 às 18h
      - Salas virtuais informadas no e-Disciplinas

roneysantos@usp.br

Dúvidas?


O que vemos aqui?

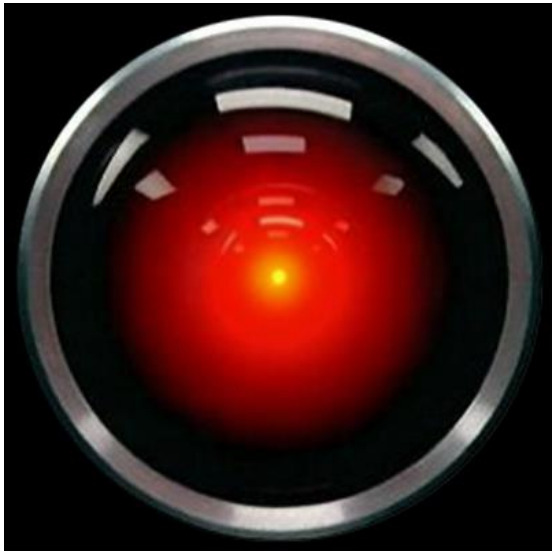
A Aurora dos Homens

HAL e Bowman

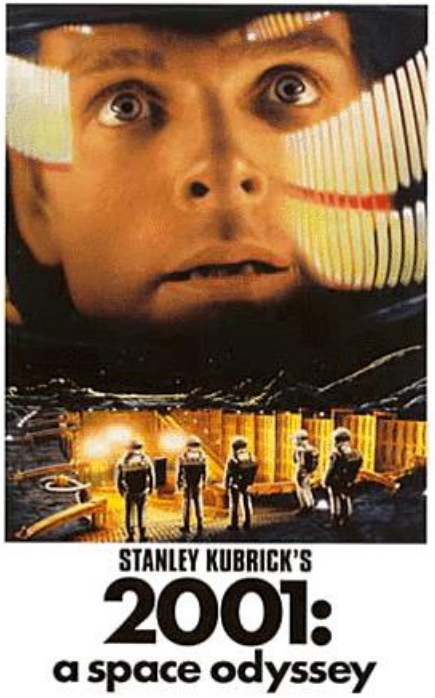
HAL vs Bowman

Meta clássica da IA e do PLN

- HAL 9.000 (**H**euristically programmed **AL**gorithmic Computer)
  - Incrível capacidade de linguagem
  - Inspiração clássica em PLN



1968



Diversas referências culturais

- Jornada nas Estrelas
- Guerra nas Estrelas
- IA
- Matrix
- Eu, robô
- O homem bicentenário
- Wall-E
- Ela
- Ex-Machina: Instinto Artificial
- Etc.

O quão longe estamos?

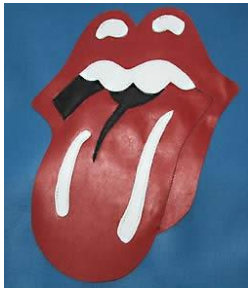
Para construir um computador como HAL

- Requer um volume enorme de conhecimento de uma dada língua
  - Reconhecimento (faz até leitura labial) e síntese de fala (fonética e fonologia)
  - Conhecimento das palavras envolvidas (morfologia e vocabulário)
    - Significado (semântica) e como combinam (uso das palavras)
  - Como grupos de palavras de juntam (gramática)
  - Manter um diálogo (discurso)
    - É educado responder... mesmo que você queira matar alguém (HAL)
    - É educado ser cooperativo... mesmo que esteja fingindo (HAL)

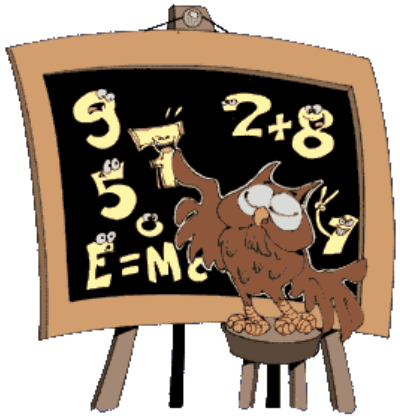
- O uso de língua natural também pressupõe conhecimento do mundo e de senso comum

Língua Natural

- Língua humana



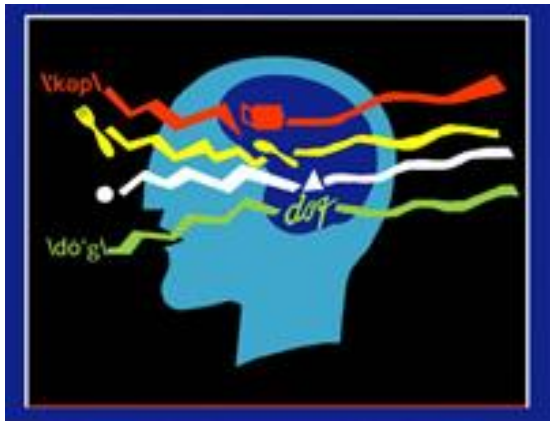
- Em oposição às linguagens artificiais
  - Matemática, lógica, linguagens de programação de computadores



PLN

- Processamento de Língua Natural
  - Linguística Computacional
  - Processamento de Linguagem Natural
  - Engenharia das Línguas Naturais

- No Brasil, tradicionalmente visto como subárea da Inteligência Artificial & Computação
  - Habilidade linguística é um tipo de inteligência



Questão

- Qual a diferença entre “língua” e “linguagem”?
  - É Processamento de Linguagem Natural ou Processamento de Línguas Naturais?

Linguagem & língua

- Linguagem: capacidade humana de comunicação e suas manifestações, de forma verbal ou não
  - Fala, gestos, música, dança, pintura, *um sorriso*
  - Envolve nosso aparato físico e mental/cognição

- Língua: código de comunicação utilizado por uma comunidade, com suas regras específicas
  - Português, Inglês, LIBRAS, etc.

PLN

- Instruir o computador a lidar com a língua, ou, como se diz informalmente, a “ler e escrever”
  - Interpretação de textos
  - Tradução automática
  - Revisão gramatical
  - Busca de respostas para perguntas
  - Sumarização
  - Auxílio a escrita e ao aprendizado de línguas
  - Etc.

- Olhares complementares
  - Computação
  - Linguística

PLN: um pouco de história

- Nascimento na 2ª guerra mundial
  - Tradução automática

- Possíveis nomes
  - *Computational Linguistics*
  - *Mechanolinguistics*
  - *Automatic Language Data Processing*
  - *Natural Language Processing*

PLN: um pouco de história

- Trajetória da Inteligência Artificial
  - Primeiros trabalhos → grande expectativa → resultados pobres → desilusão e hibernação da área → novos horizontes
    - Relatório da ALPAC (*Automatic Language Processing Advisory Committee*): *Languages and Machines – computers in translation and linguistics*

PLN: um pouco de história

- Globalização, internet, tecnologia da informação, Google

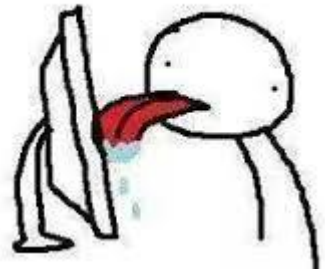
PLN: um pouco de história

- Globalização, internet, tecnologia da informação, Google
  - Mais recentemente, web 2.0, redes sociais, smartphones, *big data* e ciência de dados, *deep learning*

PLN

- Auxílio às tarefas humanas
  - (Ainda) Não substitui o humano
    - Não é possível “automatizar” toda a língua, apenas aspectos dela
    - O computador (ainda) é uma máquina estúpida!

VS.



PLN: Eliza

- Programa psicólogo dos anos 60
  - Muitas versões na Internet
- Parece muito esperto, mas se perde depois de algum tempo de conversa
  - Muitas pessoas foram enganadas pelo programa!!!
- Muito simples
  - Se baseia na fala da pessoa para formular sua pergunta/resposta

PLN: Eliza (Jackson e Moulinier, 2002)

Eliza: como vai você hoje?
P: não muito bem
Eliza: por que você diz não muito bem?
P: eu estou com alergia
Eliza: interessante
P: eu fico com manchas vermelhas
Eliza: acho que não sei do que você está falando
P: manchas vermelhas no meu rosto
Eliza: seu rosto
P: isso mesmo
Eliza: é isso mesmo?

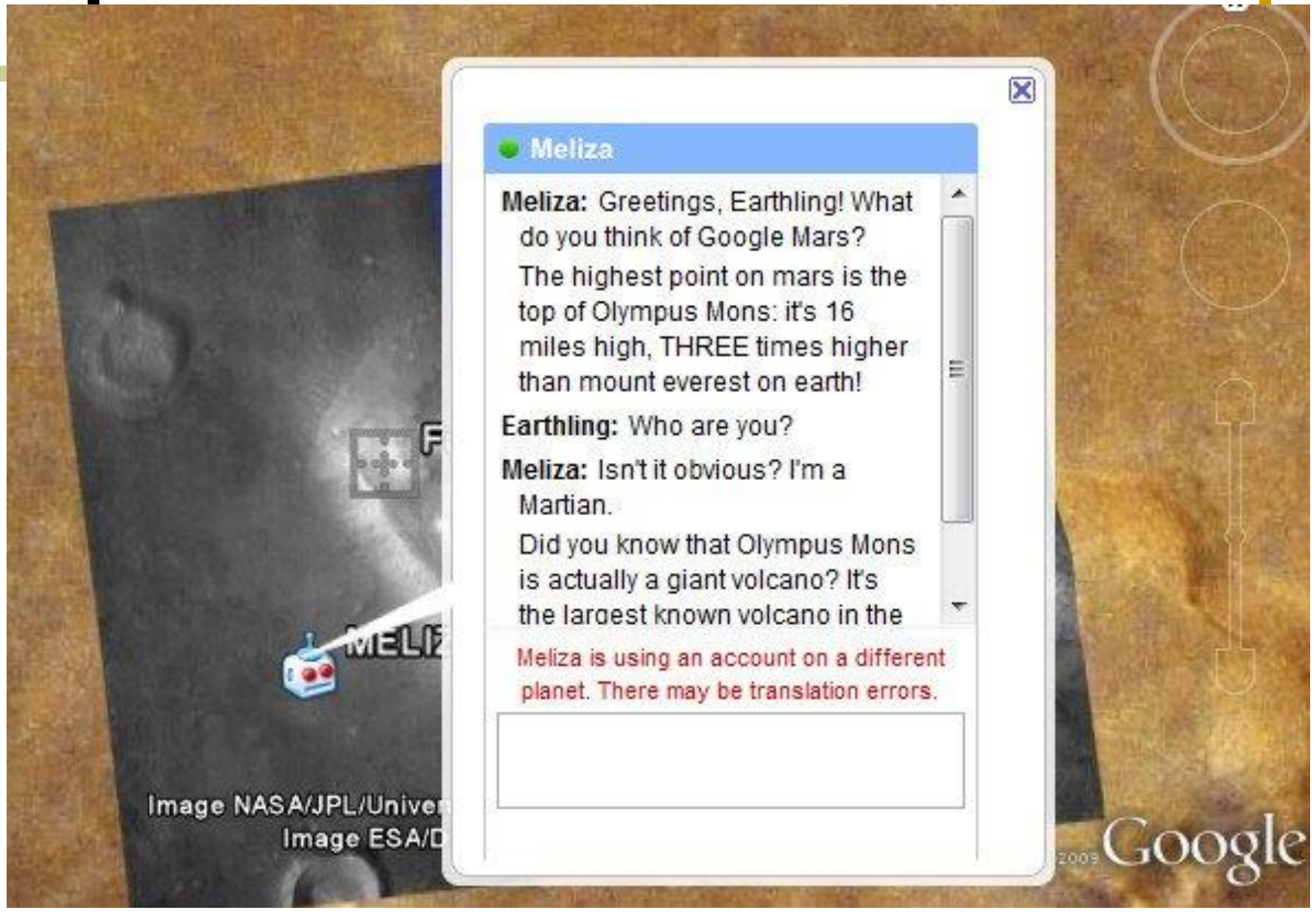
PLN: Eliza (Jackson e Moulinier, 2002)

P: você é um programa, não é?
Eliza: te agrada pensar que eu sou um programa?
P: com certeza
Eliza: eu percebo
P: ok
Eliza: um pouco antes você falou de seu rosto
P: e sobre o seu rosto?
Eliza: o que é que você realmente quer saber?
P: você é meu tipo de garota?

PLN: Eliza (Jackson e Moulinier, 2002)

Eliza: talvez em suas fantasias eu seja seu tipo de garota
P: em minhas fantasias você é loira
Eliza: o que te faz pensar que sou loira?
P: não sei
Eliza: você realmente não sabe?
P: não
Eliza: você está dizendo não apenas para ser negativo?
P: não
Eliza: você está sendo um pouco negativo
P: não, não estou
Eliza: por quanto tempo você não tem estado?

Eliza e Google Mars



PLN

- “Conversar” com uma máquina não é tão difícil
- Fazer a máquina “entender” é difícil, talvez impossível
  - Futuro distante
    - Muitas pesquisas
      - Programas especializados
      - Recursos linguísticos e linguístico-computacionais

PLN

- Será que é preciso entender de fato a língua para fazer coisas fantásticas?

WolframAlpha

WolframAlpha computational knowledge engine

dolar

Examples Random

Assuming "dolar" is a unit | Use as a character or a city instead

Assuming US dollars for "dolar" | Use Australian dollars or more ▾ instead

Input interpretation:

$1 (US dollar)

Local currency conversion for $1 (US dollar):

R$3.68 (Brazilian reais) (at current quoted rate)

Exchange history for $1 (US dollar):

Reverse rate | Last year ▾

1-year minimum	R$2.92 (27.04.2015 | 10 months ago)
1-year maximum	R$4.18 (23.09.2015 | 5 months ago)
1-year average	R$3.51 (annualized volatility: 18%)

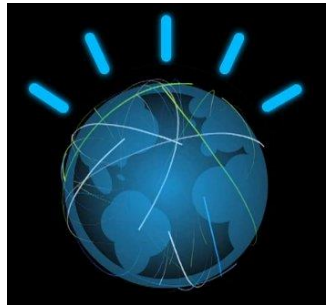
Units »

START

- *Natural Language Question Answering System*
  - http://start.csail.mit.edu/index.php

Watson (IBM)

- Venceu os melhores participantes humanos no show de perguntas e respostas Jeopardy!
  - *“more than 100 different techniques are used to analyze natural language, identify sources, find and generate hypotheses, find and score evidence, and merge and rank hypotheses”*
  - *“sources of information include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson also used databases, taxonomies, and ontologies. Specifically, DBPedia, WordNet, and Yago were used”*



Siri (Apple)

GPT

- https://talktotransformer.com/

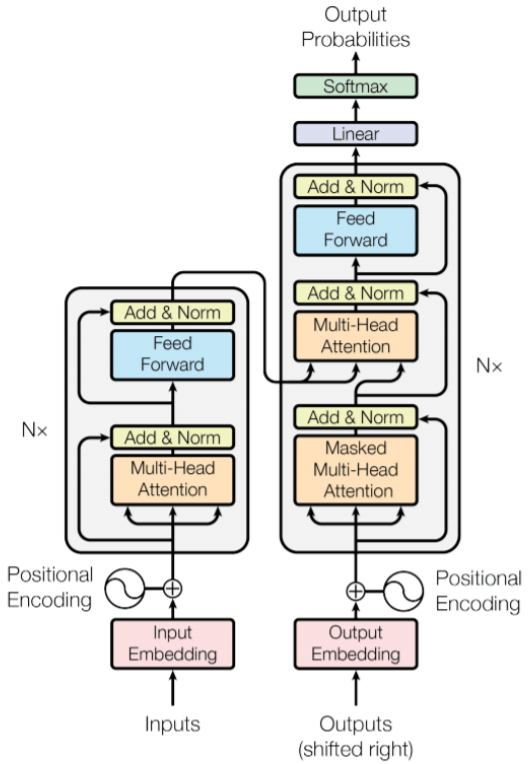


Figure 1: The Transformer - model architecture.

PLN

- Será que é preciso ser tão fantástico para ser útil?
  - Exemplos de programas simples que são úteis?

PLN

- Será que é preciso ser tão fantástico para ser útil?
  - Sugestão de possíveis sinônimos
  - Revisão ortográfica e gramatical
  - Outros?
    - Simples? Ou mais claros e facilmente automatizáveis?

PLN

- Será que é preciso ser tão fantástico para ser útil?
  - Comando wc do UNIX é um programa de PLN? Usa conhecimento linguístico?
    - Definição
      - *Short for "word count", wc displays the total number of bytes, words and lines in a text file*
      - Exemplo (em linha de comando)

```
C:\> wc meu_arquivo.txt
5 (linhas) 13 (palavras) 57 (bytes/caracteres)
```

Prática

LX-Suite

- Abrir LX-Suite, na página do grupo LX-Center (http://lxcenter.di.fc.ul.pt/) e testar as ferramentas abaixo
  - Syllabifier
  - Verbal Lemmatizer
  - POS Tagger
  - Constituency Parser
  - Dependency Parser
  - Named Entity Recognizer
  - Semantic Role Labeller
  - Semantic Similarity

- Questões
  - O que cada ferramenta faz? Alguma cometeu erro?

PALAVRAS

- Um dos parsers mais utilizados para a língua portuguesa

  - *Eckhard Bick (born 16 July 1958) is a German-born Esperantist who studied medicine in Bonn but now works as a researcher in computational linguistics. He was active in an Esperanto youth group in Bonn and in the Germana Esperanto-Junularo, a nationwide Esperanto youth federation. Since his marriage to a Danish woman he and his family live in Denmark. (Wikipedia)*

  - http://beta.visl.sdu.dk/visl/pt/

OpenWordnet-PT

- Uma das várias possibilidades para o português
  - http://wn.mybluemix.net/

Para casa

Atividades

- Leitura da semana
  - *A língua portuguesa na era digital*
    - Disponível no e-Disciplinas

- Provinha 1
  - Deve ser feita até o início da próxima aula