
Introdução à PLN na Área da Saúde

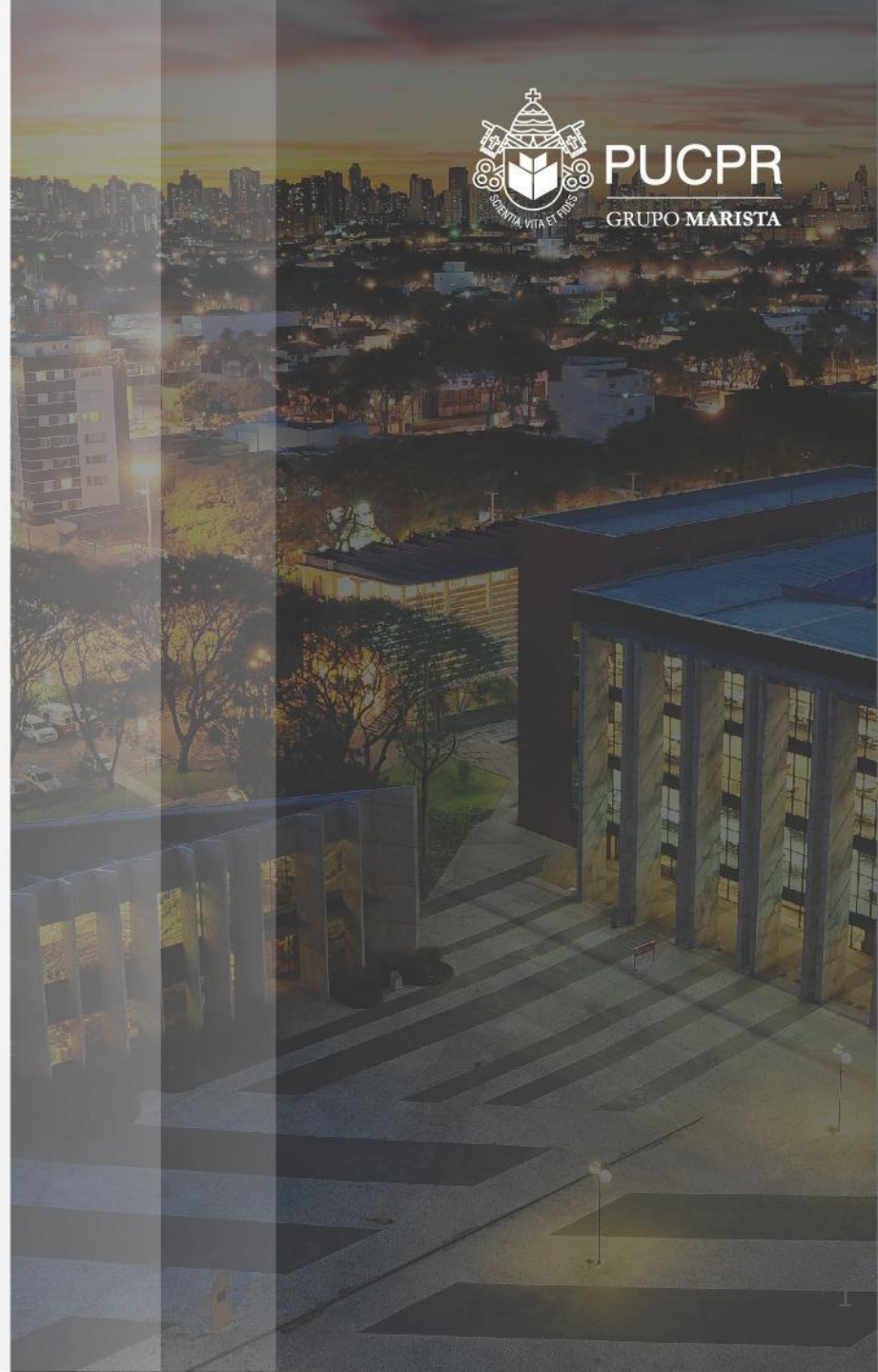
Elisa Terumi Rubel Schneider



Health Artificial Intelligence Lab



PUCPR
GRUPO MARISTA



Agenda

- Apresentação
- Processamento de Linguagem Natural (PLN)
 - Aplicações
 - Pré-processamento
 - *Word embeddings*
- Aplicações na área da Saúde
 - Desafios
 - Iniciativas HAILab
- Mãos à obra

The background of the slide is a dark, moody photograph of a workspace. It features a laptop, a pair of glasses, and a pen resting on a desk. Overlaid on this image is a faint, light-colored geometric pattern consisting of various shapes like triangles and hexagons. A thin red vertical line is positioned to the left of the title.

Apresentação

Sobre mim

- Elisa Terumi Rubel Schneider
- Mestre em Bioinformática (UFPR)
- Doutoranda em Informática (PUCPR)
- Analista e desenvolvedora de *software* (Celepar)

<https://www.linkedin.com/in/elisa-terumi-rubel-schneider/>

<https://github.com/lisaterumi>

HAILab - Health Artificial Intelligence Lab

- Iniciativa da PUCPR - PPGTS
- Grupo de pesquisa multidisciplinar
- Soluções para saúde utilizando PLN e ML
- Parcerias nacionais e internacionais



<https://github.com/HAILab-PUCPR/>

<https://huggingface.co/pucpr>

<https://www.linkedin.com/company/hailab-pucpr/>

Brasileiras em PLN

- Grupo de mulheres brasileiras que trabalham com PLN
- Dar visibilidade ao trabalho desenvolvido para a língua portuguesa
- Disseminar conhecimento para grupos de interesse
- Estimular a inserção e permanência de mulheres nessa área.



<https://brasileiraspln.com/>

<https://www.youtube.com/@brasileiraspln>

<https://www.linkedin.com/company/brasileiras-em-pln>

Exemplo de narrativa clínica

- Quais os sintomas do paciente?
- Qual tratamento está sendo realizado?
- Qual o (possível) diagnóstico? CID?
- Quando o paciente precisa receber medicação?

HAS HÁ MAIS DE 10 ANOS
NEGA DM

DRC ESTAGIO IV - POSSIVEL NEFROPATIA HIPERTENSIVA

CR: 2.1 TFG 17 ML/MIN COM ALBUMINA DE 429MG E CALCIO IONICO DE 1.44
US DE AP. URINARIO: (23/05/14)

RIM DIREITO COM COM VOLUME E ESPESSURA CORTICOMEDULAR
DISCRETAMENTE REDUZIDOS

NAO HA EVIDENCIA DE DILATACAO DO SISTEMA
COLETOR, AUSENCIA DE IMAGEM DE LITIASE

RIM MEDINDO 8,6X3,8X4,2

RIM ESQUERDO COM VOLUME E ESPESSURA CORTICOMEDULAR DISCRETAMENTE
REDUZIDOS.

NAO HA EVIDENCIA DE DILACATAO.

Em uso: ANLODIPINO 5mg, SINVASTATINA 20 mg, FUROSEMIDA 60MG/DIA,
LOSARTANA 50 MG/DIA, ALOPURINOL 100 MG/DIA

DRC POR NEFROESCLEROSE HIPERTENSIVA BENIGNA COM TFG ESTIMADA EM 24

Qual o esforço necessário para um humano extrair essas informações?

O que vamos aprender?

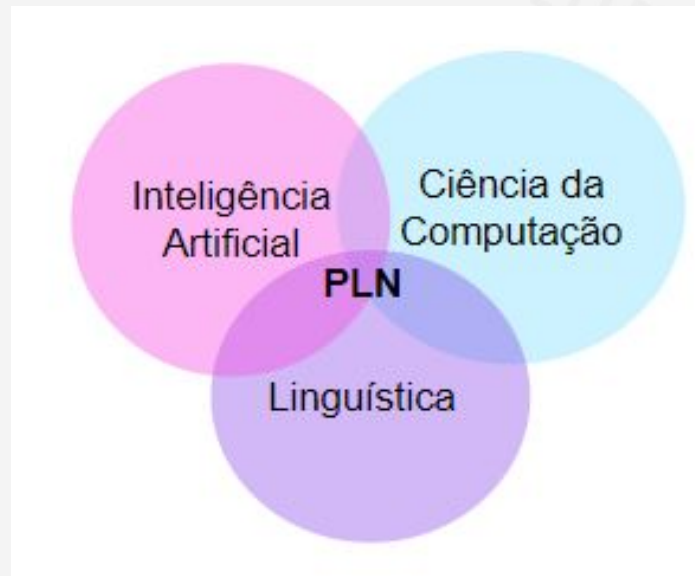
- Conceitos básicos de PLN
 - Para que serve?
 - Histórico
 - Aplicações gerais
- Aplicações na área da saúde
 - Desafios
- Mãos à obra:
 - Obter um *corpus* clínico
 - Tokenização, *stemming*, *stopwords*
 - Reconhecimento de entidades
 - Word2Vec

The background of the slide is a dark, semi-transparent image of a desk. On the desk, there is a laptop, a pair of glasses, and a pen. The text 'Processamento de Linguagem Natural' is overlaid on the image in a white, sans-serif font. A vertical red line is positioned to the left of the text.

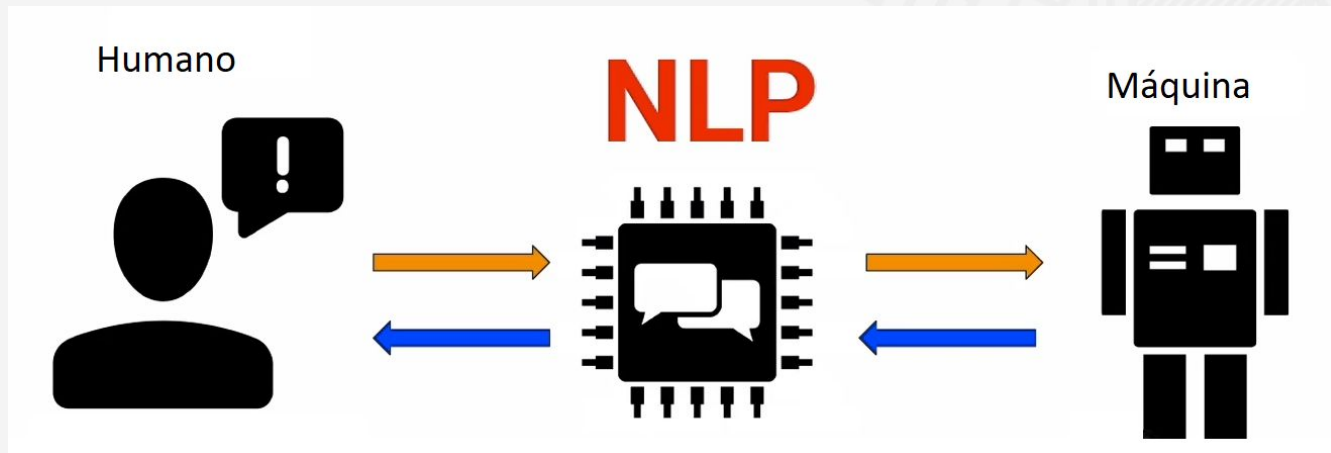
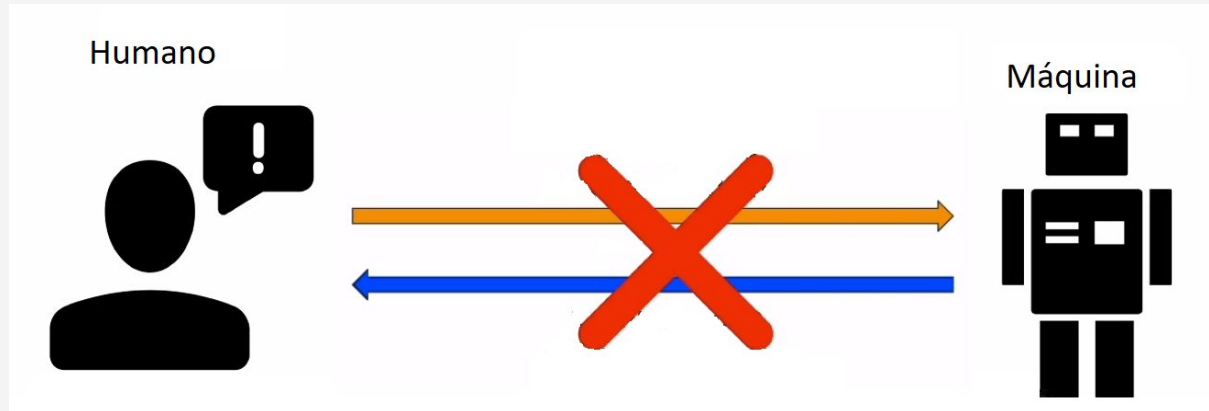
Processamento de Linguagem Natural

Processamento de Linguagem Natural

- Área interdisciplinar
- Ajuda computadores a entender, interpretar e manipular a linguagem natural (humana)
- Linguagem escrita ou falada

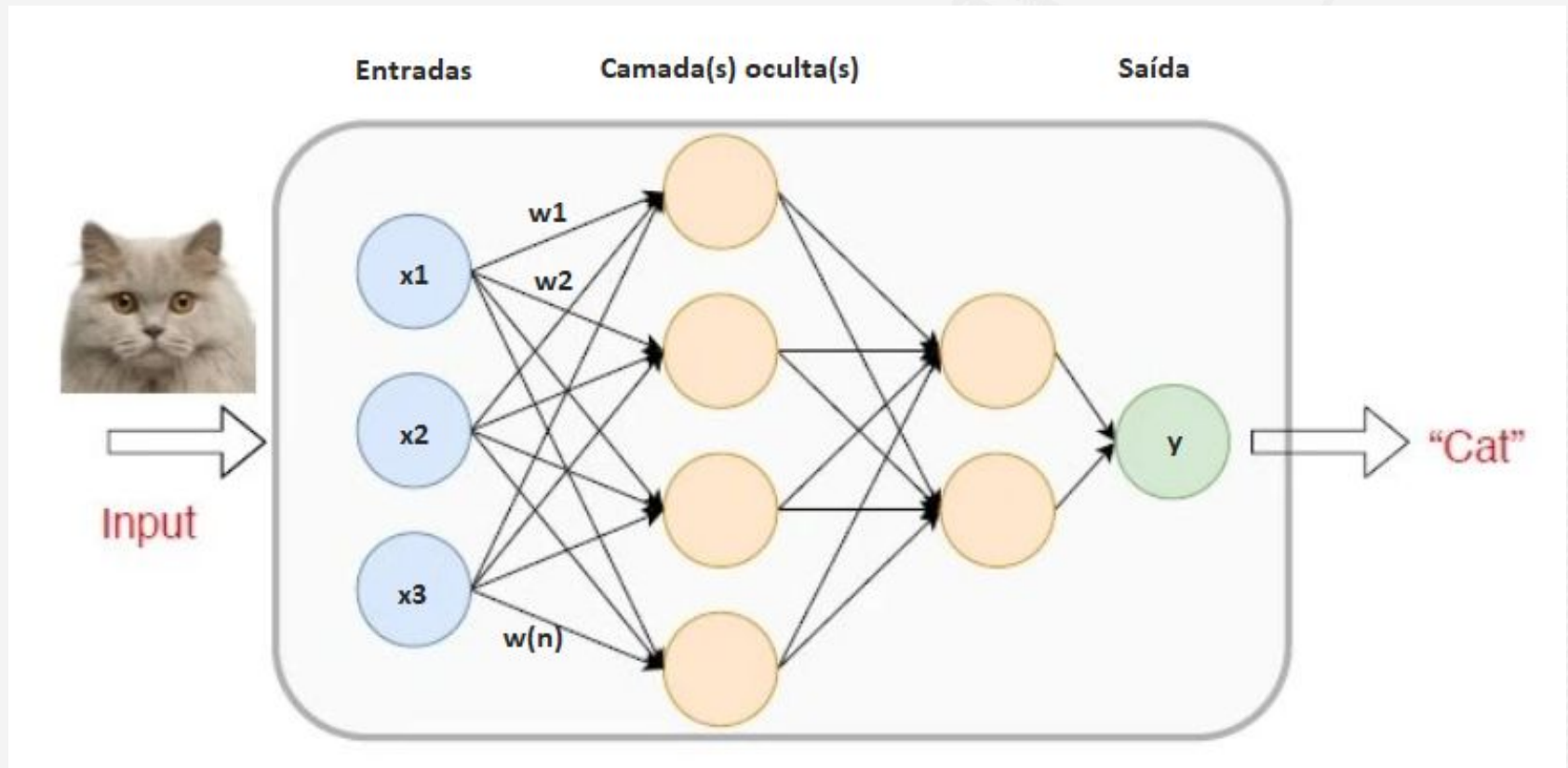


Processamento de Linguagem Natural



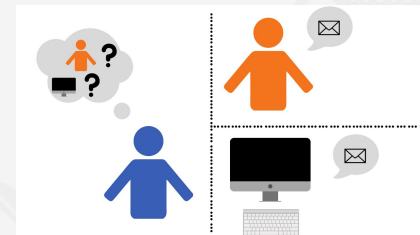
Processamento de Linguagem Natural

Rede neural



Processamento de Linguagem Natural

Histórico



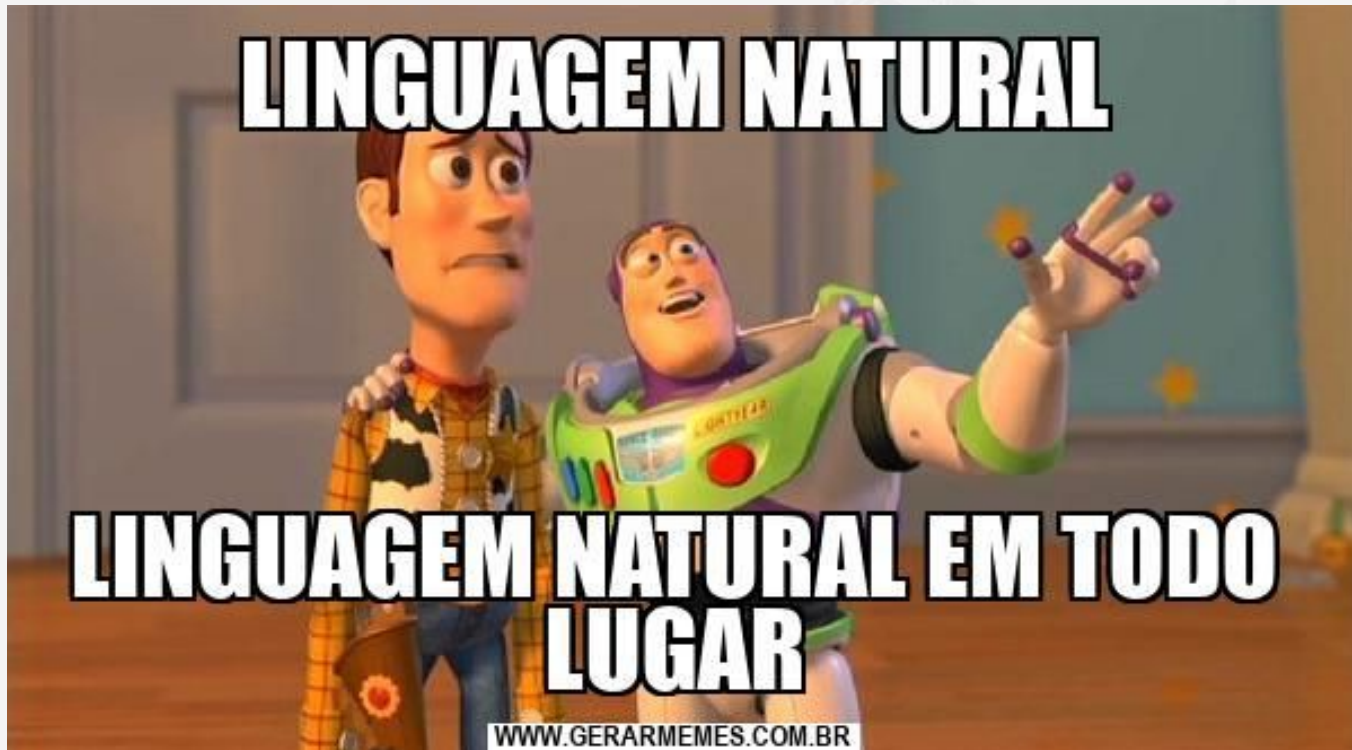
- Teste de Turing (1950)
 - Está conversando com uma máquina ou uma pessoa?
 - Perguntas e respostas de uma conversa comum
- 1960: [ELIZA](#)
 - Simula a conversação entre um humano e um computador
 - “Impressão” de que entende o que o humano fala
- 1980: sistemas baseados em regras começaram a proliferar
 - Aprendizado de Máquina (*Machine Learning*)
 - Grande passo para a evolução da área

Porque o interesse em PLN?

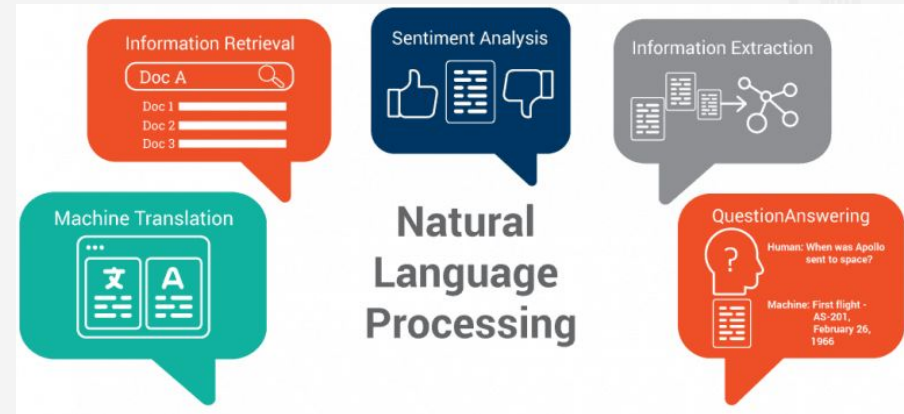
- Há muita informação textual (dado não estruturado) acumulada na *Web*, nas empresas, nos computadores das pessoas.

The Amazon logo, featuring the word "amazon" in a black, lowercase, sans-serif font, with a curved orange arrow underneath it pointing from the letter 'a' to the letter 'z'.The Twitter logo, consisting of the word "twitter" in a white, lowercase, sans-serif font, followed by a white bird icon, all on a blue rectangular background.The YouTube logo, featuring the word "You" in a black, sans-serif font, followed by the word "Tube" in a white, sans-serif font inside a red rounded rectangle.The Google logo, with the word "Google" in its characteristic multi-colored font (blue, red, yellow, blue, green, red) on a white background.The WhatsApp logo, featuring a green speech bubble with a white telephone handset inside, followed by the word "WhatsApp" in a black, sans-serif font.The Facebook logo, with the word "facebook" in a white, lowercase, sans-serif font on a blue rectangular background.

Porque o interesse em PLN?



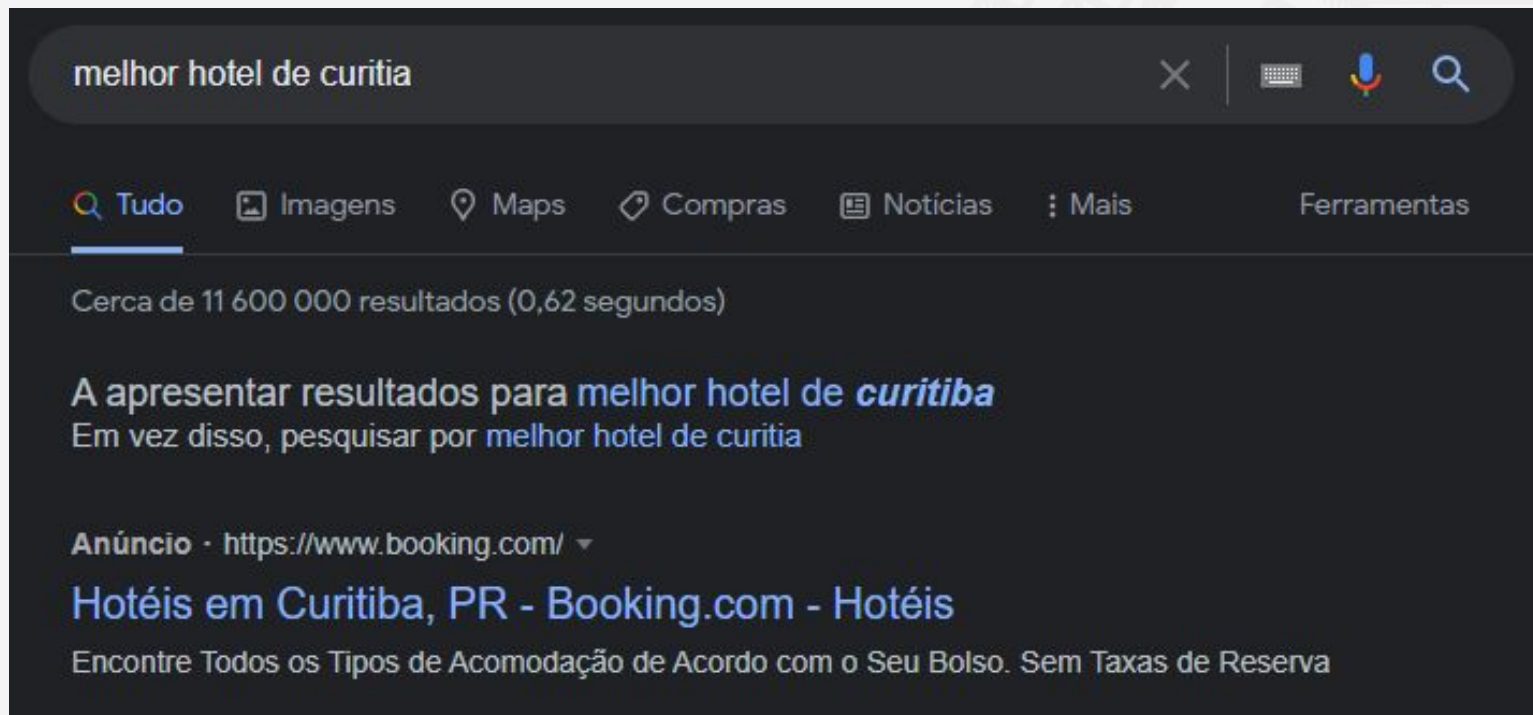
Aplicações de PLN



- Das mais simples:
 - Busca por palavra-chave
 - Identificação de sinônimos
 - Verificação da escrita (ortografia)
 - Extração da informação
- Às mais sofisticadas:
 - Tradução automática
 - Reconhecimento e geração da fala
 - Geração de textos sintéticos
 - Sistemas de diálogo e Chatbots

Exemplos

- Recuperação da Informação

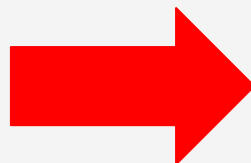


Exemplos

- Extração da informação a partir de dados textuais

Olá Cris, gostaria de agendar uma reunião sobre o projeto Identidade Visual. Estarei na Rua XV de Novembro, 1145 dia 15 de janeiro das 10:00h as 11:30h.
-Pedro

Criar evento no Calendário



Evento: Reunião
Projeto: Identidade Visual
Data: 15/01
Início: 10:00h
Fim: 11:30h
Local: Rua XV de Novembro, 1145

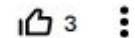
Exemplos

- Análise de Sentimentos



Anônimo

Rio das Ostras, RJ • 8 contribuições



Passeio

dez de 2019

A viagem de trem foi muito boa, mas poderia ter sido melhor se "nos momentos das atrações" a velocidade fosse reduzida permitindo um melhor registro fotográfico.

A van, no retorno, apesar de ser uma Mercedes, apresentou espaço de poltrona apertado, causando um pouco de desconforto.

O guia foi excelente, muito simpático e cordial.

Excelente passeio, recomendo.

Exemplos

- Tradução

The screenshot displays the Google Translate web interface. At the top, the Google Tradutor logo is visible. Below it, there are tabs for 'Texto', 'Documentos', and 'Sites'. The language selection bar shows 'PORTUGUÊS' as the source language and 'JAPONÊS' as the target language. The input text on the left is 'o que é o processamento de linguagem natural?'. The translated text on the right is '自然言語処理とは何ですか?'. Below the Japanese text, the romanized version 'Shizen gengo shori to wa nanidesu ka?' is shown. The interface includes icons for voice input/output, a character count (45 / 5.000), and a feedback link at the bottom right.

Google Tradutor

Texto Documentos Sites

DETECTAR IDIOMA PORTUGUÊS INGLÊS ESPANHOL ▼

↔ JAPONÊS ESPANHOL INGLÊS ▼

o que é o processamento de linguagem natural? ×

自然言語処理とは何ですか? ☆

Shizen gengo shori to wa nanidesu ka?

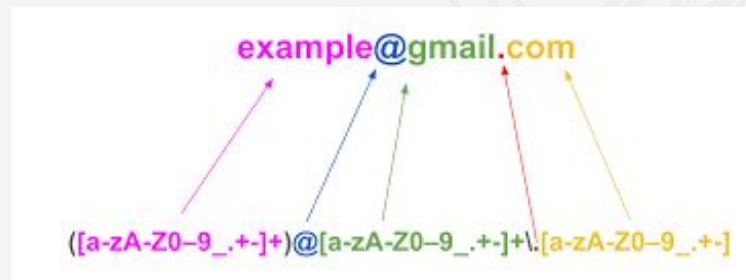
45 / 5.000

Enviar feedback

Operações básicas em textos

Pré-processamento

- Tokenização
- Normalização
- POS tagger
- Lematização
- *Stemming*
- Retirada das *stopwords*
- Expressões regulares



Word embeddings

- Representar palavras em vetores numéricos
 - Computadores trabalham com números
- Técnicas de representar palavras
 - *One-hot encoding*
 - Matriz Tf-idf
 - *Word embeddings* vetoriais (ex word2vec)
 - *Word embeddings* contextuais (ex BERT)



Word embeddings

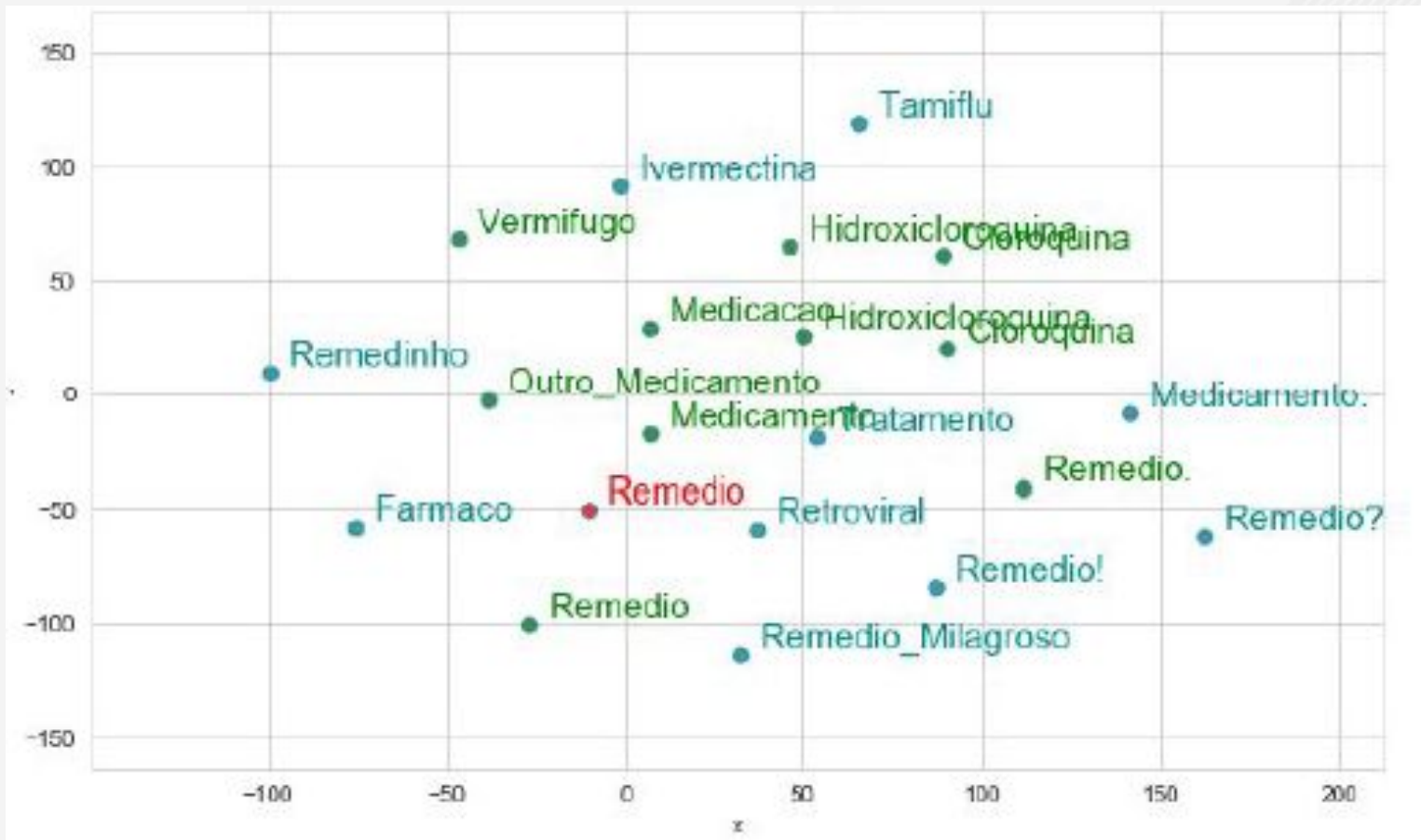
Word2Vec

- Tomas Mikolov et al (Google), 2013
- Modelo preditivo que recebe um *corpus* como entrada e retorna vetores de características
- Cada vetor carrega significado semântico da palavra
- Semelhança semântica e sintática e relação com outras palavras, com base na forma em que as palavras são usadas no texto
- Permite operações aritméticas

Para conhecer melhor: <http://jalammar.github.io/illustrated-word2vec/>

Word embeddings

Word2Vec



[W2V com tweets brasileiros sobre Covid-19 treinado por um aluno de iniciação científica](#)



Aplicações na área da Saúde

Registro Eletrônico em Saúde (RES)

- Sistema informatizado que armazena informações de saúde de pacientes
- Narrativas clínicas
- Histórico médico
- Resultados de exames
- Prescrições, alergias, vacinações, etc

Grande volume de textos clínicos e prontuários eletrônicos contendo informações valiosas dos pacientes.

Melhoria no cuidado ao paciente

- Apoio à decisão clínica
 - Predição de complicações pós-cirúrgicas
 - Alerta de progressão de doenças
 - Sumarização de dados relevantes do paciente
 - Prevenção de readmissão do paciente



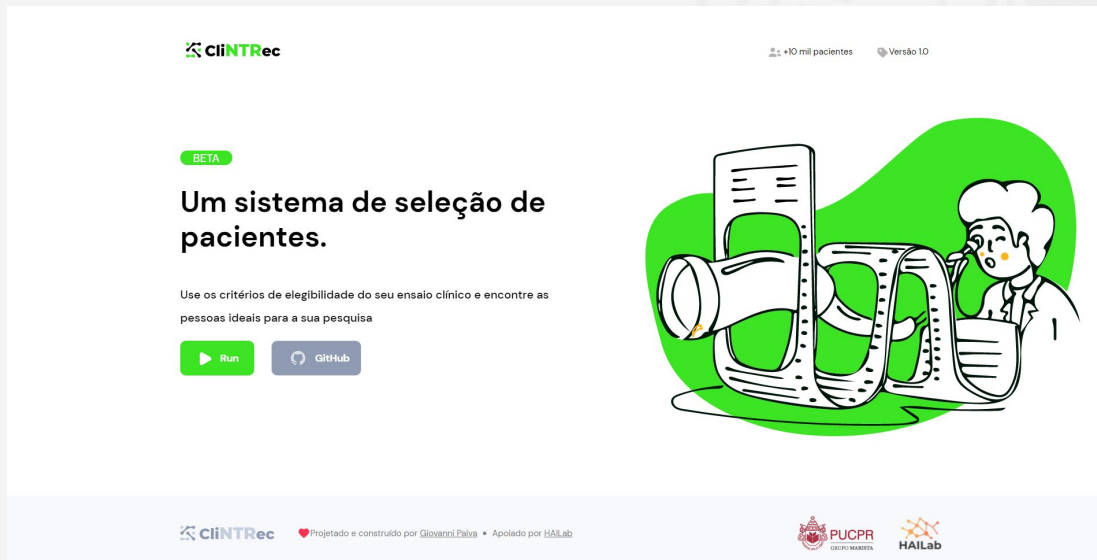
Otimização de processos médico-administrativos

- Redução de custos
 - Cobrança eficiente: Atribuição de CID automática
 - Controle de erros administrativos: Evitar repetição de processo e atrasos em procedimentos
- Eficiência administrativa
 - Chatbots (atendimentos e triagem)
 - Reconhecimento de fala para agilizar documentação



Apoio a indústria farmacêutica

- *Screening* de pacientes
 - Recrutamento de participantes para ensaios clínicos
- Farmacovigilância
 - Monitoramento de eventos adversos a medicamentos



The screenshot displays the CliNTRec website. At the top left is the CliNTRec logo. To the right, it indicates '+10 mil pacientes' and 'Versão 1.0'. Below the logo, a green 'BETA' badge is present. The main heading reads 'Um sistema de seleção de pacientes.' followed by the text 'Use os critérios de elegibilidade do seu ensaio clínico e encontre as pessoas ideais para a sua pesquisa'. There are two buttons: a green 'Run' button with a play icon and a grey 'GitHub' button with the GitHub logo. On the right side, there is a cartoon illustration of a doctor in a white coat and stethoscope, looking at a large, stylized green and white graphic that resembles a medical device or a large 'E'. The footer contains the CliNTRec logo, a heart icon with the text 'Projetado e construído por Giovanni Palva • Apoiado por HAILab', and logos for PUCPR (PUC do Rio de Janeiro) and HAILab.

Desafios

- Falta de estrutura formal
- Especulação e negação
- Erros ortográficos e gramaticais
- Falta de acentuação e pontuação
- Abreviação (ambíguas) e acrônimo
- Informações implícitas
- Informações contraditórias
- Heterogeneidade
 - Tamanho do texto
 - Formato de escrita
 - ...

Paciente encaminhado da unidade básica devido inapetência e fraqueza, com histórico de anemia. Em exame anterior observado sangue oculto nas fezes.
Conduta: realizado pedido de colonoscopia e retorno com exames.

DRC DE ETIOLOGIA INDEFINIDA
PARCIAL NORMAL
ECO COM CISTO RENAL SIMPLES A DIREITA
CREATININA DE 2.5 COM CLEARENCE ESTIMADO DE 24 ML/MIN
PA: 110/70
K: 5.8
ATR? HIPOALDO?
SOLICITO RENINA ALDOSTERONA E GASO VENOSA
RETORNO COM EXAMES EM 3 MESES

#DRC ESTAGIO 4 DE CAUSA INDETERMINADA
#SEM COMORBIDADES
#NEGA TABAGISMO
#NEGA ETILISMO

#S: SEM QUEIXAS NO MOMENTO. RETORNA COM EXAMES

#O: BEG, LOTE, HIPOCORADO, HIDRATADO
PA:140/80 FC:83
BCRNF RCR 2T MV+ SRA
SEM EDEMA DE MMII

EXAMES:
US 04/06/13: RIM D: 9 X 4 X 3 RIM E: 9 X 5X 3
CISTO RENAL SIMPLES À DIREITA

LAB:28/05/15:
MICROALBUMINURIA: 19.2 CKDEPI: 21ML/MIN
Cr: 2.9 Ur: 59 Hb: 11,3 Ht: 34,1 Na: 139 K: 6,2 P: 3,8

Desafios

Narrativa clínica

ID : Lucas, 79 anos, viuvo, cobrador de ônibus aposentado, 6 filhos. # Medicções em uso : - losartana 50 mg 12/12 h - AAS 100 mg - captopril ? 1 x ao dia -hidroclorotiazida 25 mg 1 x ao dia -flutamida 250 mg 8/8 h # Comorbidades : HAS, Ca de Prostata em acompanhamento há 2 anos. Nega tontura, dispneia, dor no peito, cansaço, ou outras queixas desde a consulta anterior.

≠

Texto jornalístico

A possibilidade de a jornada de trabalho ser composta por apenas quatro dias da semana será testada no Brasil a partir de novembro. Após ser experimentado no Reino Unido, Estados Unidos, Espanha, Austrália e outros países, o modelo será colocado em prática no Brasil por meio de uma parceria entre a consultoria Reconnect Happiness e a organização 4 Day Week Global.

≠

Wikipedia

Cesare Mansueto Giulio Lattes, mais conhecido como César Lattes (Curitiba, 11 de julho de 1924 — Campinas, 8 de março de 2005), foi um físico brasileiro, descobridor do méson- π (méson pi ou pión), descoberta que levou à concessão do Prêmio Nobel de Física de 1950 a Cecil Frank Powell, líder da pesquisa.

Desafios

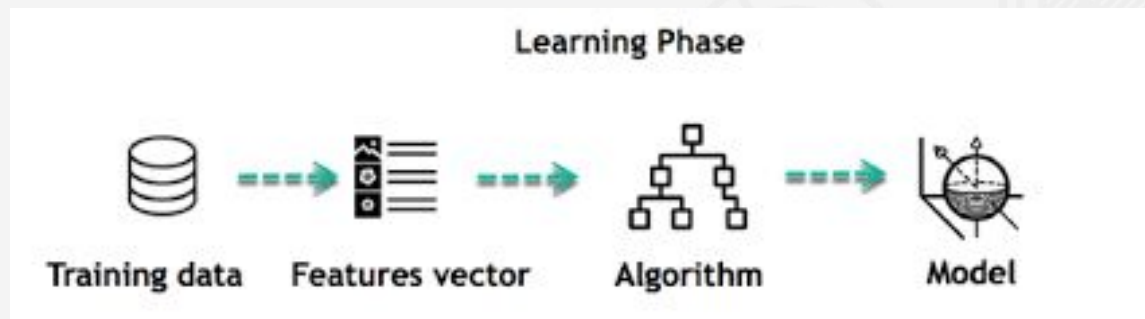
- Falta de recursos semântico-linguísticos
 - Recursos escassos para o domínio clínico em Português
 - Dificuldade de acesso aos dados
 - Custo de de-identificação e anotação de dados (rotulação)
 - Necessidade de infra (GPU) para treinar modelos robustos.



Iniciativas HAILab

Corpora (*datasets*)

- SemClinBr
- SummClinBr
- MorphClinBr
- TempClinBr
- DepClinBr (*in progress*)
- NestedClinBr (*in progress*)



Aprendizado supervisionado: necessita dados anotados (rotulados)

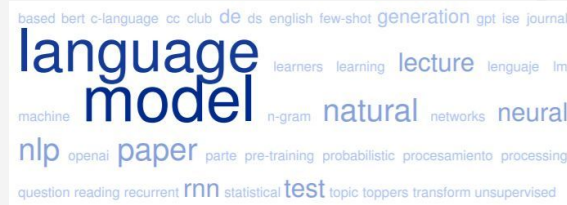
Corpus NestedClinBr

- (a) Problema
Tabagista de 78 maços / ano por 65 anos (cigarro de palha).
- (b) Problema
sem queixas. sente-se bem.
- (c) Problema Problema
Relata o aparecimento de angina há mais ou menos 3 meses, aos mínimos esforços.
- (d) Problema Anatomia
internamento por edema agudo de pulmão, última sendo em Agosto

Token	Classe
Tabagista	Problema
de	O
78	O
maços	O
/	O
ano	O
por	O
65	O
anos	O
(O
cigarro	O
de	O
palha	O
)	O
.	O

Exemplo de anotação de entidades clínicas (interface e arquivo texto)

Modelos de Linguagem



- Word2Vec com *tweets* sobre Covid-19
- GPT-2-Biomédico
- BioBERTpt
- CardioBERTpt
- *Question-Answering* (QA) clínico/biomédico
- Modelos para NER com Flair
- Modelos para NER com CRF
- Modelos para NER com BERT
- Modelos de sumarização
- Modelos para extração de relação temporal
- (...)

Leitura Recomendada:

- [Curso de NLP - HuggingFaces](#)
 - [NLP e saúde: como a IA pode promover transformações no setor](#)
 - [NLP in Healthcare Trends and Challenges in 2022](#)
 - [NLP for Machine Learning in Healthcare](#)
 - [PLN e Cuidados de Saúde](#)
 - [Benefícios dos Chatbots de Serviços de Saúde](#)
 - [Tecnologias de inteligência artificial utilizadas na saúde durante a pandemia de covid-19](#)
-

The background is a dark, moody photograph of a workspace. It includes a laptop keyboard, a pair of glasses, and a hand-drawn diagram of a mechanical or structural component. A hand is visible, holding a pen and pointing at the diagram. The overall tone is professional and focused.

Mãos à obra!

Google Colab



- Serviço de nuvem gratuito hospedado pela Google
- Acesso pelo navegador
- Versão Colab PRO: R\$ 58,00
- Incentivar pesquisa em Aprendizado de Máquina e IA
- Formato “[notebook](#)”:
 - Código-fonte (geralmente em python)
 - Texto rico (*markdown*) com imagens
 - Resultado do códigos
- Ambiente colaborativo -> compartilhar com colegas

[\[PRATICA\]IntroducaoPLNSaude_2023.ipynb - Colaboratory \(google.com\)](#)

Obrigada

<https://www.linkedin.com/in/elisa-terumi-rubel-schneider/>



PUCPR
GRUPO MARISTA

