



Ciclo de  
Encontros:  
4 x PLN

## Cap. 22 Extração de Informação 3a ed

imagem

**Daniela Barreiro Claro**

Joaquim Santos

**Marlo Souza**

**Renata Vieira**

Glória Pinheiro

11/03/2024

# Sumário

---

- ❑ Um pouco de história
- ❑ Conceituação Formal
- ❑ Extração de Informação | REN e EIA
- ❑ Avaliação
- ❑ Considerações Finais

## Extração de Informação

é a tarefa de obter informação estruturada de dados não estruturados.





Documentos



Sentença

“Renata Costa palestrou na Universidade Federal da Bahia”

Extração de Informação

Entidades

“Renata Costa”

“Universidade Federal da Bahia”

Relações

“palestrou”

## Pré-Processamento

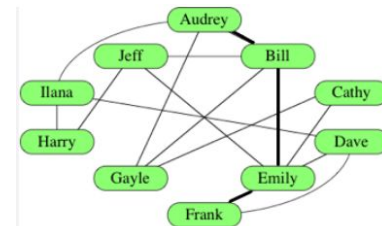
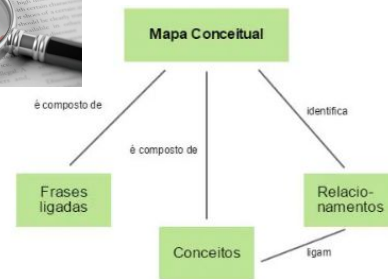
## Extração de Informação

Entidades

Relações

Entidades

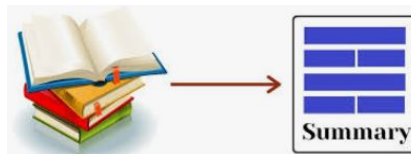
Relações



## Sentiment Analysis



Positive Negative Neutral



EI

Própria tarefa

Extração de Informação

Entidades

Relações

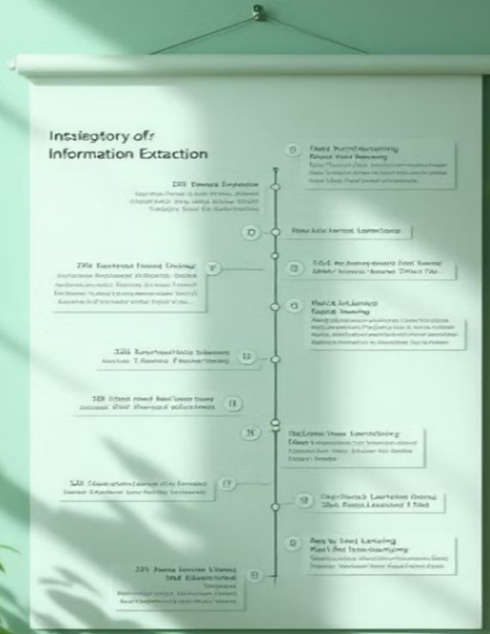
Entidades

Relações

MÉTODOS

Sentença →





# História

Primeiros trabalhos com gramáticas formais e parsers sintáticos.

Década de 1970

Métodos de aprendizado de máquina e semi-supervisionado.

Década de 2000

Década de 1990

Conferências MUC impulsionam o desenvolvimento da área.

Atualmente

Redes neurais e grandes modelos de linguagem.



# Eventos Notáveis

---

## **MUC (1987 - 1997)**

Foco em extração de templates a partir de textos militares e jornalísticos. A MUC-6 introduz a tarefa de REN, o que foi responsável pelo grande avanço e amadurecimento dos métodos para essa tarefa

## **ACE (1999 - 2008)**

Sucede a MUC mantendo objetivos similares, porém com configurações próximas às tarefas tais quais as conhecemos hoje. Distingue claramente REN, ER e EE como subtarefas da EI. Introdução de outras línguas (chinês e árabe)

## **HAREM (2004-2008)**

Conferências de Avaliação Conjunta para REN (e uma forma limitada de ER) para a língua portuguesa



# Conceituação



# Entidade e Relação

---

## Entidade

Objeto concreto ou abstrato, como pessoa, livro, sentimento. (Chen, 1976).

Tradicionalmente entendido como designadores rígidos, como nomes próprios, referências temporais e valores.

## Relação

- Associação entre entidades (Chen, 1976)
- Relacionamentos entre entidades que atuam como veridadores de alguma proposição relacionando duas ou mais entidades, ou seja, uma relação mantida entre essas entidades (Guarino, Guizzardi, 2015)

## “a é uma maçã”

A existência de um objeto denotado pelo nome a que é uma maçã é uma condição suficiente para a verdade desta frase.



# Entidade e Relação

---

## No contexto de EI

Relações são como tipos de relacionamentos de primeira ou segunda ordem

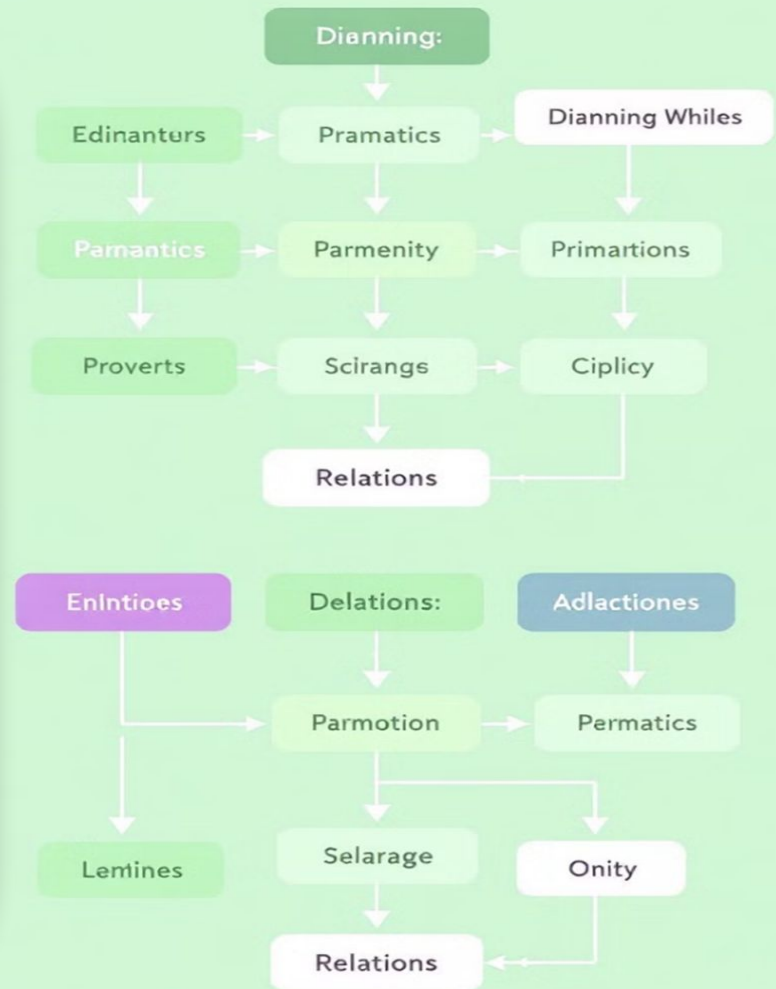
### “Filósofos são autores de Livros”

- Relações descrevem ou implicam propriedades de classes gerais.
- É um tipo de relacionamento que existe entre objetos, suas propriedades e classes de objetos ou suas propriedades.

A noção de “relação” e “entidade” na literatura da área mais recente requer mais aprofundamento por demandar um significado diferente principalmente por diferentes visões de autores.



# Extração de Informação





Sentença:

**Renata Costa palestrou na Universidade Federal da Bahia em Salvador.**

**EI Tradicional**

**Relação específica**

**Extrações**

**palestrar\_em**

**palestrar-em** (Renata, Salvador)

**EI Aberta**

**Extrações**

(Renata, **palestrou em**, Salvador)  
(Renata, **palestrou em**, a Universidade Federal da Bahia)

# Extração de Informação Tradicional



## Objetivo

Extrair informações estruturadas de tipos pré-definidos.



## Limitação

Necessidade de intervenção humana para especificar novos fatos.

| Relação Específica              | Exemplo de Sentença                    | Extração                   |
|---------------------------------|----------------------------------------|----------------------------|
| location-of(algo/alguém, local) | Um aluno pode ser encontrado na escola | location-of(aluno, escola) |
| is-a(subclasse, superclasse)    | Salvador é uma cidade                  | is-a(Salvador, cidade)     |
| part-of(todo, parte)            | Roda é um componente de um carro       | part-of(roda, carro)       |

# Reconhecimento de Entidades Nomeadas (REN)

Entities:

Beginner Terenities

Person.

Organization

Green irzatina sion. "New am "Lmce"

For Location



Sentença:

**Renata Costa** palestrou em **Salvador**.

Renata Costa PER palestrou em Salvador LOC .

É a tarefa de identificar e classificar expressões linguísticas, denominadas entidades nomeadas (EN) e que referenciam entidades específicas no domínio do discurso, como nomes próprios, expressões temporais e espécies biológicas.



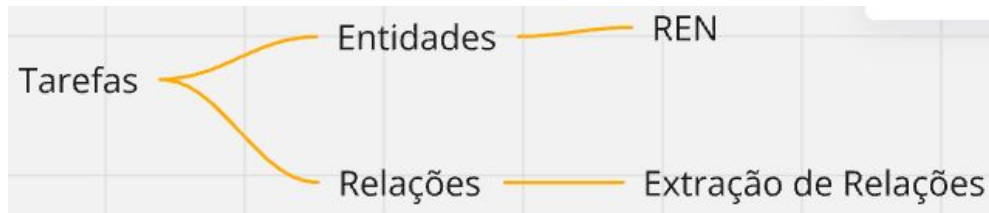
## **Identificação da expressão**

As palavras que formam a EN são identificadas.



## **Classificação**

É atribuída uma categoria semântica à EN.



Sentença:

**Amanda Silva** palestrou na **Universidade Federal da Bahia** em **Salvador**.



## Identificação da expressão

“Amanda Silva”

“Universidade Federal da Bahia”

“Salvador”



## Classificação

**Pessoa**

**Organização**

**Local**

REN é uma  
tarefa de  
grande  
importância  
para PLN

# Métodos

Métodos ricos em conhecimento: regras e recursos linguísticos

Métodos baseado em dados, como aprendizado de máquina e arquiteturas neurais

Abordagens iniciais

Abordagens Recentes

Regras Iniciais e consultas a gazetteers

Construção automática ou semi-automática de gazetteers

Métodos baseados em classificação e classificação sequencial

# Métodos



## Classificação Sequencial

Esquema BIO - B (begin); I (inside); O (outside)

Amanda/**B-PER**

Silva/**I-PER**

palestrou/**O**

na/**O**

Universidade/**B-ORG**

Federal/**I-ORG**

da/**I-ORG**

Bahia/**I-ORG**



## Modelos Neurais Recentes

BART (Lewis et al., 2020a),  
RoBERTa (Liu et al., 2019b),  
T5 (Raffel et al., 2020),  
BERT (Devlin et al., 2019)  
RELP-CRF (Amaral; Vieira,  
2014)

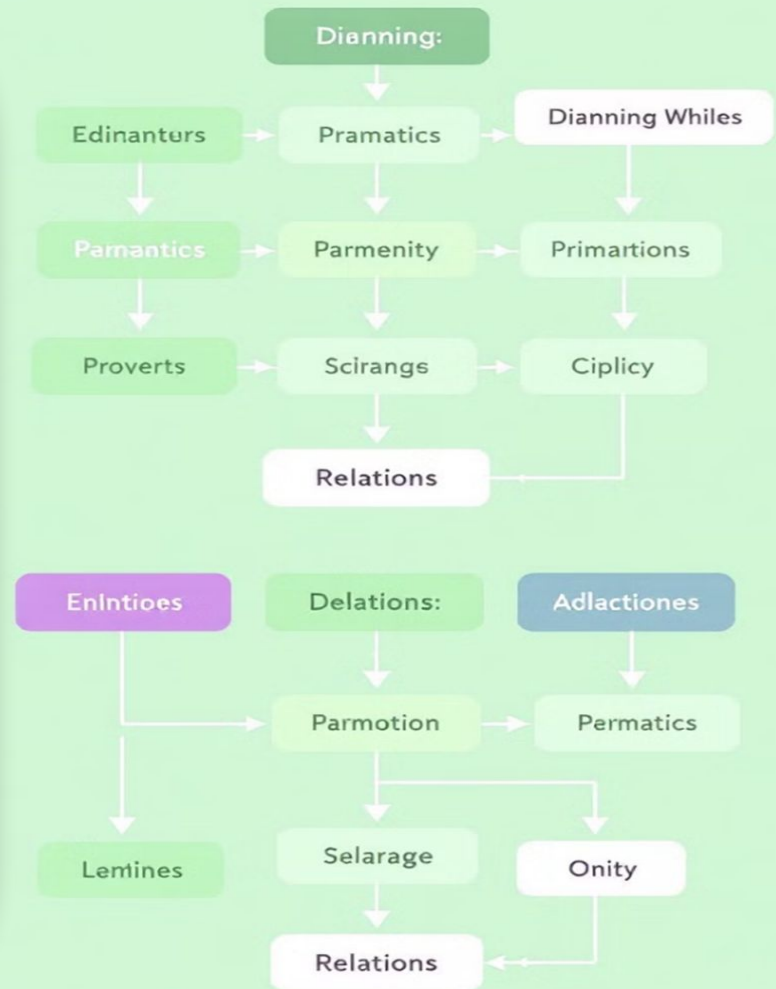
# Domínios

---

## Domínios Diversos

O reconhecimento de entidades tem sido aplicado em muitas áreas específicas, como direito, saúde, biologia, geologia e história.

# Extração de Relações



# Extração de Relações

É a tarefa de identificar os relacionamentos entre entidades de um determinado escopo mencionadas em um texto. O escopo refere-se a um conjunto de relações alvo de um determinado domínio de conhecimento ou aplicação.

| Relação                              | Descrição                                                                                                                                                               | Exemplo                                 |
|--------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------|
| Pertence(Cidade, Unidade Federativa) | Sobre uma <b>cidade</b> que está localizada em uma determinada <b>Unidade Federativa</b> , dizemos que a primeira <i>pertence</i> a esta última.                        | Pertence(Salvador, Bahia)               |
| Tem_Prefeito(Cidade, Pessoa)         | Uma <b>pessoa</b> que executa a função administrativa de gestão do executivo em nível municipal de uma dada <b>cidade</b> é denominada de seu(sua) <i>prefeito(a)</i> . | Tem_Prefeito(Salvador, Bruno Reis)      |
| Fundação(Cidade, Data)               | A <b>data</b> em que uma <b>cidade</b> foi fundada, é dita sua data de <i>fundação</i> .                                                                                | Fundação(Salvador, 29 de março de 1549) |

# Extração Conjunta de Entidades e Relações

As tarefas de entidades nomeadas e extração de relações são interdependentes.

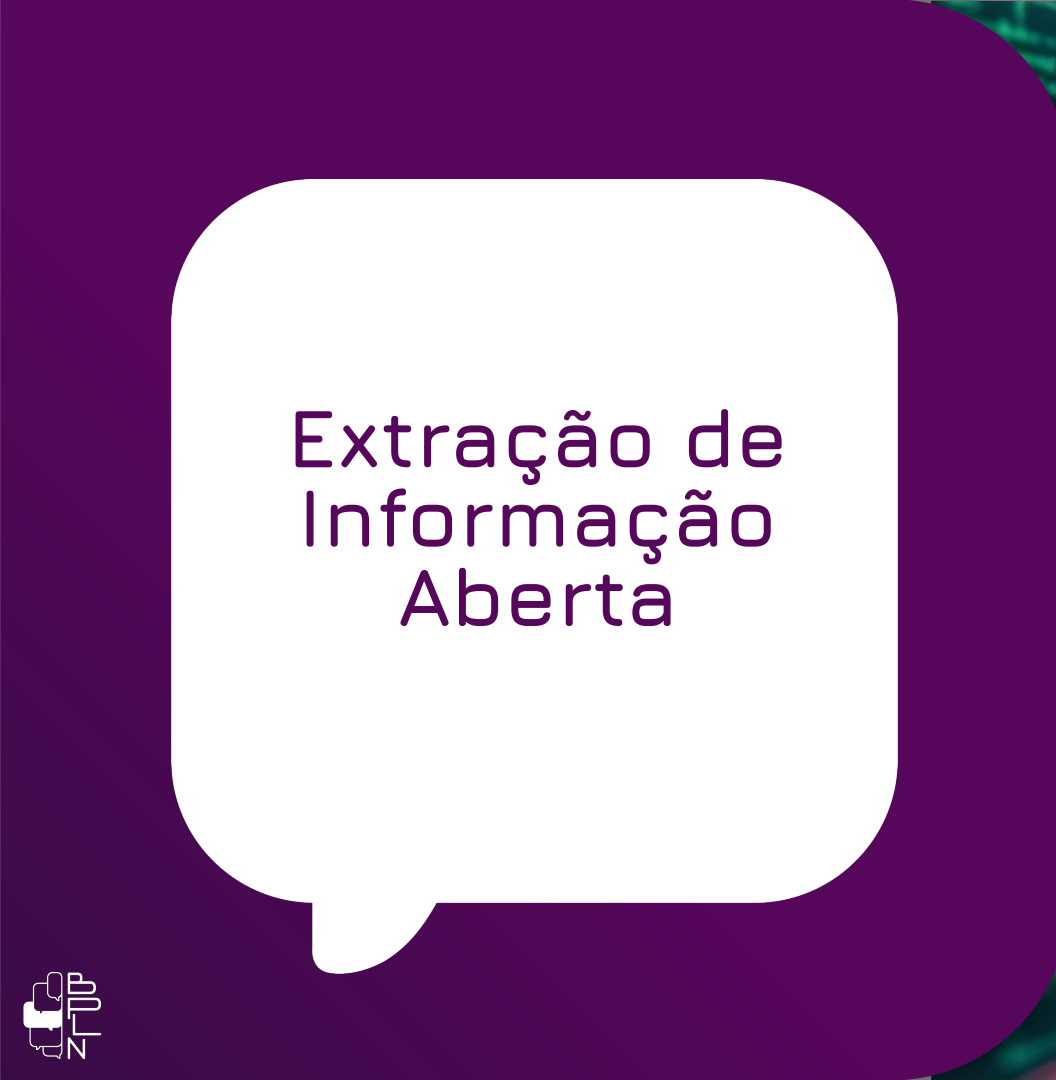
O escopo delimita tanto as categorias e a natureza quanto as relações entre essas entidades.

Atualmente diversos trabalhos da literatura consideram a tarefa de **extração conjunta** das relações e entidades (**ERE- Entity and Relation Joint**).

Há também aqueles que fazem a extração sequencial (**REN+ER**).

A generalização da tarefa de ER para domínios abertos sem pré-definição das relações originou outra tarefa.

Extração de  
Informação  
Aberta



# Extração de Informação Aberta (EIA ou OpenIE)

Inicialmente proposta por Banko 2007, através do sistema TextRunner (Banko 2007)

Surge como uma generalização da tarefa de Extração de Relações e Entidades sem classificação prévia

Extrair informação estruturada de documentos sem pré-definir as relações e nem os tipos de entidades

# EIA para Língua Portuguesa

---

## Frase

A empresa XYZ está localizada em Salvador.

O teclado faz parte do notebook.

## Extrações

(A empresa XYZ; está localizada em; Salvador)

(O teclado; faz parte de; o notebook)

Como observado em (Glauber, Claro, 2018), a EIA para Língua Portuguesa tem uma história bastante recente.

# Conceituação Formal

Considerando uma sentença  $X$  composta de tokens  $x_i$ ,  $X = \langle x_1, x_2, \dots, x_n \rangle$

Um extrator EIA é uma função que mapeia  $X$  em um conjunto  $Y = \langle y_1, y_2, \dots, y_j \rangle$  como um conjunto de tuplas  $y_i = \langle rel_i, arg1_i, arg2_i, \dots, arg_{n_i} \rangle$ ,

Consideramos que as tuplas estão sempre no formato  $y = (arg1, rel, arg2)$ , na qual  $arg1$  e  $arg2$  são sintagmas nominais, não necessariamente formados por tokens presentes em  $X$ , e  $rel$  é um descritor de um relacionamento entre  $arg1$  e  $arg2$



from DptOIE-Web



# História

**Regras e Léxicos  
Pos-Tagger e Chunkers**

2012 - 2018

**Métodos de aprendizado de  
máquina (BIO-inspired)**

2020-2022

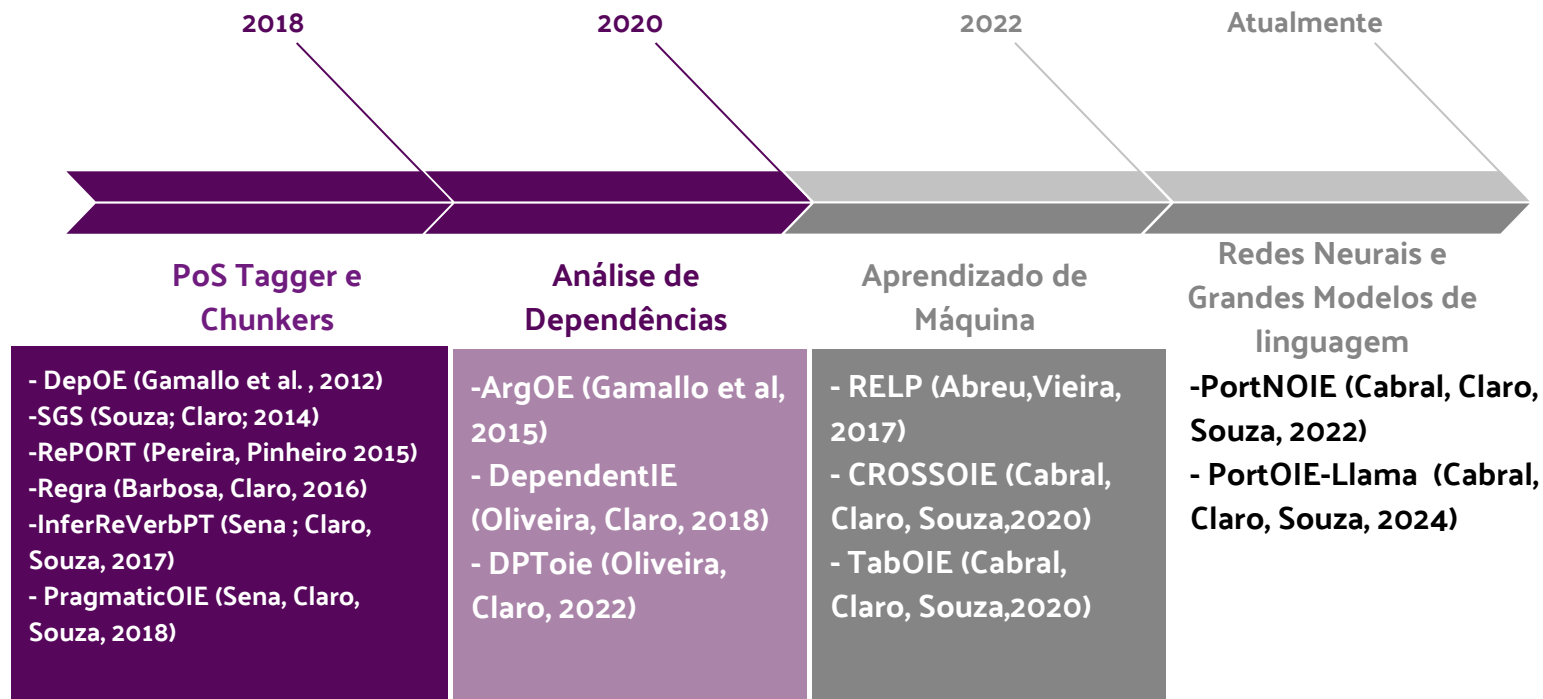
2018-2020

**Analisadores de  
Dependências**

Atualmente

**Redes neurais e grandes  
modelos de linguagem**

# Métodos



# Métodos

## PoS Tagger e Chunkers



| EXTRAÇÕES RePort (texto em português) |                           |                        |
|---------------------------------------|---------------------------|------------------------|
| <i>arg1</i>                           | <i>rel</i>                | <i>arg2</i>            |
| O Movitae                             | foi criado em             | 2003                   |
| o Congresso Nacional                  | iniciava os debates sobre | clonagem terapêutica   |
| Técnica                               | ficou fora de             | a Lei de Biossegurança |

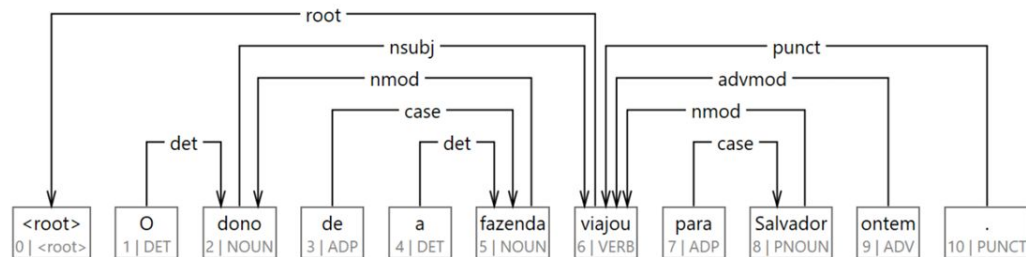
REL  $\rightarrow$  V I V P I V W \* P

RePORT (Pinheiro et al., 2015)

# Métodos

## Análise de Dependências

O dono da fazenda viajou para Salvador ontem.



Arg1

Rel

Arg2

O dono de a fazenda

viajou

para Salvador

O dono de a fazenda

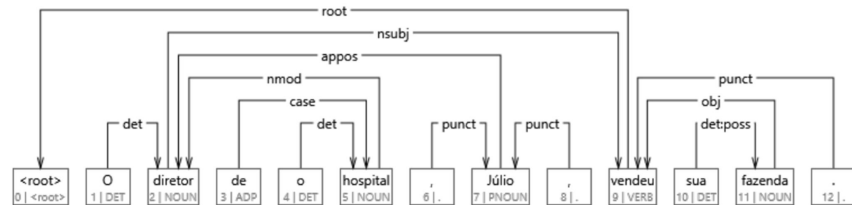
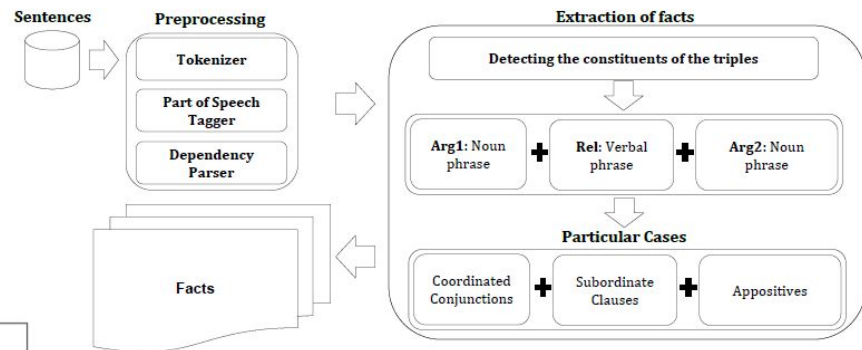
viajou

ontem

O dono de a fazenda

viajou

para Salvador ontem



| Arg1                    | Rel    | Arg2        |
|-------------------------|--------|-------------|
| O diretor de o hospital | vendeu | sua fazenda |
| O diretor de o hospital | é      | Júlio       |
| Júlio                   | vendeu | sua fazenda |

# Métodos

Aprendizado de  
Máquina



Classificador Sequencial

**O gato caçou o rato.**

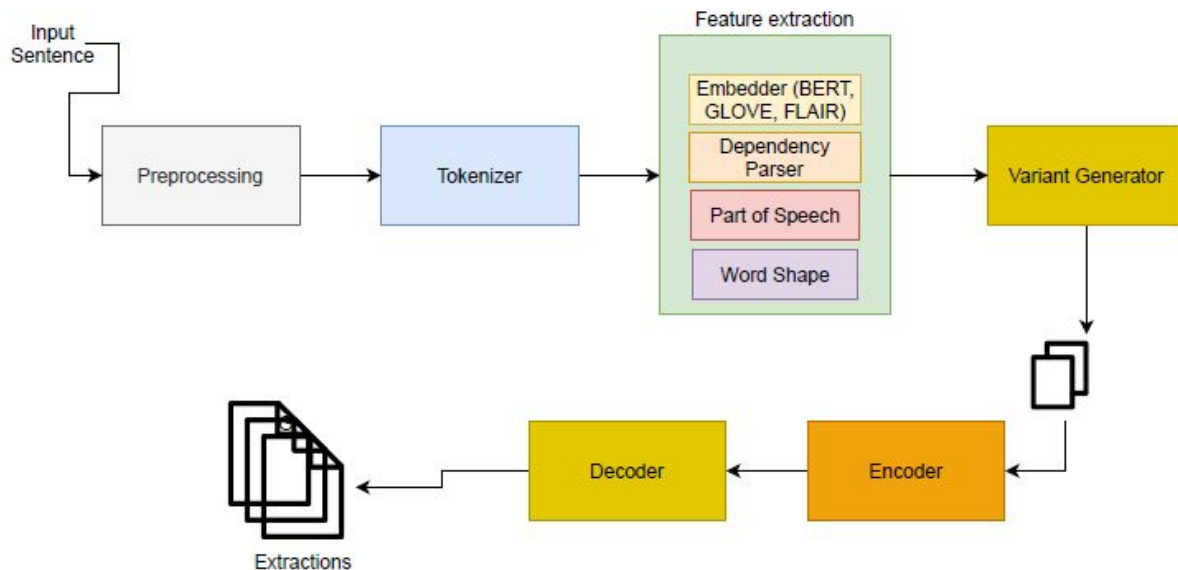
O → B-ARG1

gato → I-ARG1

caçou → PRED

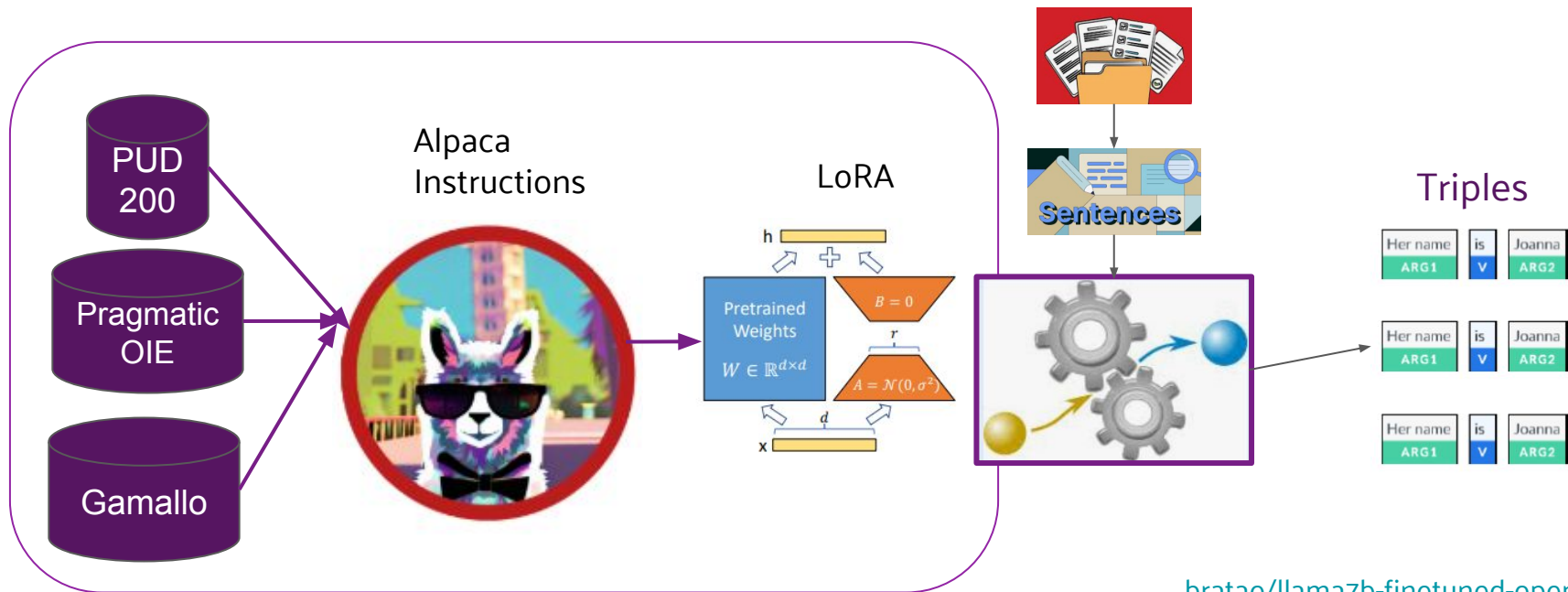
o → B-ARG2

rato → I-ARG2



# Métodos

Redes Neurais e  
Grandes Modelos de  
linguagem



[bratao/llama7b-finetuned-openie-lora](#)



Avaliação



# Avaliação

Input Sentence:

Her name 

|    |
|----|
| is |
| V  |

 Joanna and she 

|       |
|-------|
| likes |
| V     |

 to 

|       |
|-------|
| dance |
| V     |

Extraction for  
"is":

|          |
|----------|
| Her name |
| ARG1     |

|    |
|----|
| is |
| V  |

|        |
|--------|
| Joanna |
| ARG2   |

 and she likes to dance ✓

Extraction for "like":

Her name is Joanna and 

|      |
|------|
| she  |
| ARG0 |

|       |
|-------|
| likes |
| V     |

|          |
|----------|
| to dance |
| ARG1     |

 ✓

Extraction for "dance":

Her name is Joanna and 

|      |
|------|
| she  |
| ARG0 |

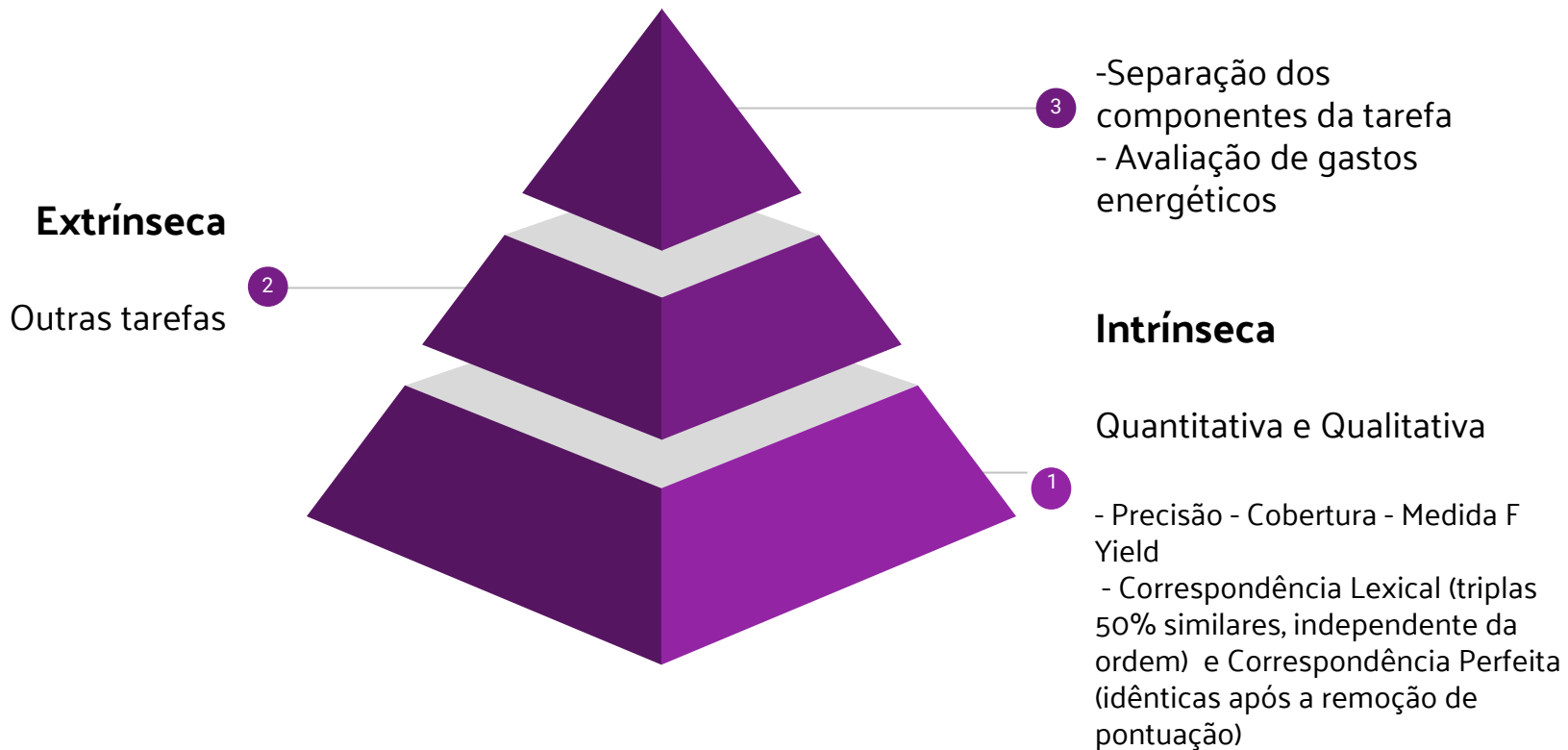
 likes to 

|       |
|-------|
| dance |
| V     |

 ✗



# Avaliação



Recursos



# Recursos REN

---

## Corpora

---

Anotados diretamente em PT

HAREM

GeoNER

ParishMemories

LENER

UlyssesNER-BR

## Domínio

---

Geral

[https://www.linguateca.pt/aval\\_conjunta/HAREM/harem\\_ing.html](https://www.linguateca.pt/aval_conjunta/HAREM/harem_ing.html)

Geologia

[GeoCorpus-V3](#)

História

<https://github.com/jneto04/ner-llm-pt/tree/main>

Legal

<https://github.com/Convenio-Camara-dos-Deputados/ulyssesner-br-propor>

Legal

<https://github.com/ulysses-camara/ulysses-ner-br>

# Recursos EIA

## Corpora

Anotados diretamente em Português

|                                     | #<br>Sentenças | #<br>Extrações |
|-------------------------------------|----------------|----------------|
| Gamallo<br>(Gamallo & Garcia, 2015) | 103            | 346            |
| Pragmático<br>(Sena & Claro, 2020)  | 400            | 485            |
| PUD 200<br>(Souza et al., 2024)     | 200            | 337            |
| PUD 100<br>(Souza et al., 2024)     | 100            | 136            |

## Ferramentas

- ArgOE
- PragmaticOIE
- DptOIE
- PortNOIE

## Semi-automáticos

- Transalign  
Transferência de uma tarefa para outra.
- Corpus Sintéticos  
Geração com LLM

<https://github.com/formas/>

[https://github.com/FORMAS/awesome\\_openie](https://github.com/FORMAS/awesome_openie)



# Desafíos



# Desafios

---



Conceituação da Tarefa



Ferramentas online



Extração de n-tuplas



Domínios



Corpora



Avaliação

# Considerações Finais



# Considerações Finais

---

Uma revisão da área de Extração de Informação

Análise das principais tarefas: NER, ER e EIA

Principais métodos e recursos utilizados nas tarefas

Desafios e problemas em aberto





Obrigada!



# Referências de Imagens

---

[Imagem 1](#)

[Imagem 2](#)

[Imagem 3](#)

[Imagem 4](#)