

---

# Extração Automática de Metainformação de Documentos

---

Prof<sup>a</sup>. Juliana Pinheiro Campos Pirovani  
Departamento de Computação  
Universidade Federal do Espírito Santo – UFES  
[juliana.campos@ufes.br](mailto:juliana.campos@ufes.br)

---

# AGENDA

- Introdução: Metainformação/ REN
  - Abordagens para Extração de Informação
  - Ferramenta CRF+LG e Resultados
  - Aplicações
-

# INTRODUÇÃO

## ■ Reconhecimento de Entidades Nomeadas (*Named Entity Recognition – NER*)

Pois, o resto, tinha grandes amigas sempre. Constante, até era, mais até que as minhas primas... tinha as minhas primas, e tinha a Joaquina Sampaio e Melo, que era filha da maior amiga da minha mãe, da Lourinhã.



Pois, o resto, tinha grandes amigas sempre. Constante, até era, mais até que as minhas primas... tinha as minhas primas, e tinha a <EM ID="H2-Efin-201" CATEG="PESSOA" TIPO="INDIVIDUAL">Joaquina Sampaio e Melo</EM>, que era filha da maior amiga da minha mãe, da <EM ID="H2-Efin-202" CATEG="LOCAL" TIPO="HUMANO" SUBTIPO="DIVISAO">Lourinhã</EM>.

# INTRODUÇÃO

## ■ Reconhecimento de Entidades Nomeadas (*Named Entity Recognition – NER*)

– Importância

– Dificuldades

E tinha a **Joaninha Sampaio e Melo**, ...

Hoje foi um ótimo dia para **Mercedes**.

... e surgem os heterónimos **H. M. F. Lecher** e ...

**Fenômeno** não participava de um coletivo ...

---

# ABORDAGENS PARA EXTRAÇÃO DE INFORMAÇÃO

- **Abordagens utilizadas no NER**
    - Linguística
    - Aprendizado de máquina
    - Híbrida
-

---

# FERRAMENTA CRF+LG

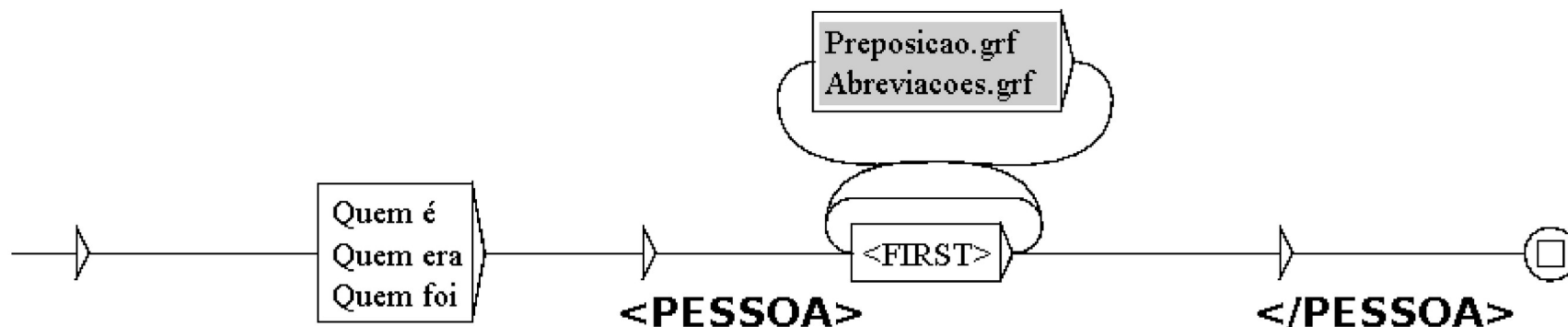
## ■ **Motivação:**

- Explorar o potencial de um sistema híbrido para o NER em Português.
    - depender de menos corpora para treino
    - usar abordagem linguística quando não há corpus de treino disponível.
  - Pesquisa em NER para o Português ainda é escassa.
-

# FERRAMENTA CRF+LG

## ■ Gramáticas Locais (LG)

- Exemplo de LGG criado no Unitex (Paumier, 2021)



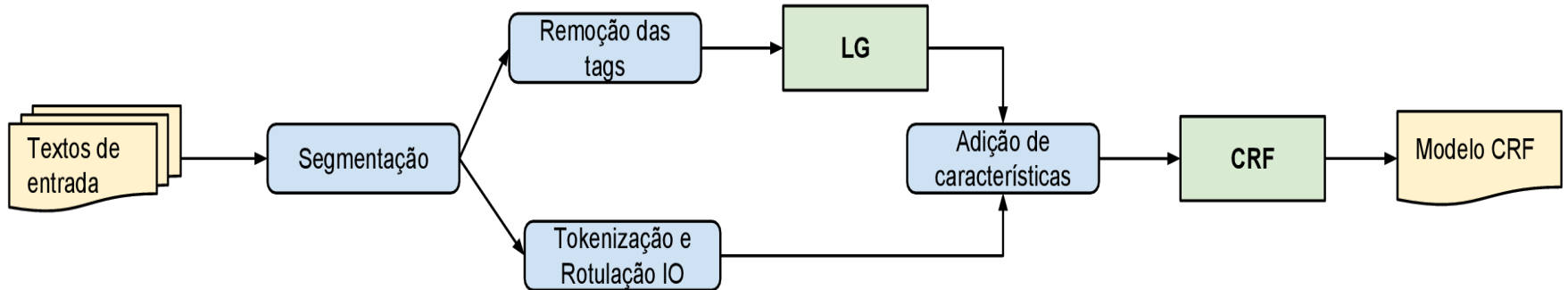
- Exemplo de concordância para o LGG apresentado

ram no atentado na Oktoberfest? </P> <P>Quem é<PESSOA> Henning Mankell</PESSOA>?{S} Como se cha  
u a construção da Torre Eiffel? </P> <P>Quem é<PESSOA> Samuel Hahnemann</PESSOA>?{S} Em homeopa  
Christian von Holst em Gdansk? </P> <P>Quem é<PESSOA> Werner Herzog</PESSOA>? </P> <P>Em que c  
é que tem lugar a Semana Verde? </P> <P>Quem era<PESSOA> Herbert Erhardt</PESSOA>?{S} Quando fo  
? </P> <P>Que significa a sigla RAF?{S} Quem era<PESSOA> Rolf Heissler</PESSOA>?{S} Diga três m

# FERRAMENTA CRF+LG

## ■ Campos Aleatórios Condicionais (CRF):

Metodologia usada



- Notação IO para a sentença “*Meu pai é Gabriel Raimundo da Silva*”: (Meu O) (pai O) (é O) (Gabriel I-PESSOA) (Raimundo I-PESSOA) (da I-PESSOA) (Silva I-PESSOA) (. O).



# FERRAMENTA CRF+LG

## ■ Métricas para avaliação de desempenho

$$\textit{Precisão} = \frac{\textit{Total de NEs identificadas corretamente}}{\textit{Total de NEs identificadas}}$$

$$\textit{Abrangência} = \frac{\textit{Total de NEs identificadas corretamente}}{\textit{Total de NEs existentes no corpus}}$$

$$\textit{Medida - F} = \frac{2 \times \textit{Precisão} \times \textit{Abrangência}}{\textit{Precisão} + \textit{Abrangência}}$$

# FERRAMENTA CRF+LG

## ■ Resultados: LG x CRF x CRF+LG

Sistemas	Identificação			Classificação		
	P (%)	A (%)	F (%)	P (%)	A (%)	F (%)
<b>LG</b>	71.27	28.48	40.70	64.80	25.06	36.14
<b>CRF</b>	79.03	66.13	72.01	64.92	52.59	58.11
<b>CRF+LG</b>	<b>79.86</b>	<b>66.76</b>	<b>72.73</b>	<b>66.52</b>	<b>53.85</b>	<b>59.52</b>

# FERRAMENTA CRF+LG

## ■ Resultados: CRF+LG x NERP-CRF (Amaral, 2013)

Sistemas	Identificação			Classificação		
	P (%)	A (%)	F (%)	P (%)	A (%)	F (%)
<b>NERP-CRF</b>	74.83	54.86	63.31	62.13	44.08	51.57
<b>CRF+LG</b>	<b>79.86</b>	<b>66.76</b>	<b>72.73</b>	<b>66.52</b>	<b>53.85</b>	<b>59.52</b>

## ■ Resultados: CRF+LG x CharWNN (Santos & Guimarães, 2015)

Sistemas	P (%)	A (%)	F (%)
<b>CharWNN</b>	65.21	52.27	58.03
<b>CRF+LG</b>	<b>67.09</b>	<b>54.85</b>	<b>60.36</b>

---

# FERRAMENTA CRF+LG

- **Resultados:** Avaliação em outros corpus (aTribuna e SIGARRA)
    - Algumas adaptações feitas na LG para um domínio específico proporcionaram ganhos significativos na Medida-F.
    - Na ausência de corpus do mesmo domínio para treino, pode ser melhor adaptar a LG do que usar um modelo treinado em outro domínio.
-

# FERRAMENTA CRF+LG

## ■ Resultados: Estudo dos limites do CRF

Experimentos	Identificação			Classificação		
	P (%)	A (%)	F (%)	P (%)	A (%)	F (%)
<b>CRF</b>	79.03	66.13	72.01	64.92	52.59	58.11
<b>CRF+LG</b>	79.86	66.76	72.73	66.52	53.85	59.52
<b>Limite inferior</b>	78.78	65.12	71.30	64.66	51.74	57.48
<b>Limite</b>	91.83	87.06	89.39	72.92	67.03	69.85
<b>Limite Superior</b>	<b>95.68</b>	<b>94.56</b>	<b>95.12</b>	<b>95.67</b>	<b>91.69</b>	<b>93.63</b>

O limite superior nos permite prever o ganho máximo que podemos alcançar combinando o CRF com uma LG melhorada ou outros classificadores.

# FERRAMENTA CRF+LG

## ■ Resultados: Competição IberLEF (2019)

Dataset	Class	P(%)	R(%)	F(%)
Police	PER	29.59	58.41	39.28
Clinical	PER	14.29	10.09	11.83
	<b>Overall</b>	56.26	56.66	56.46
General	ORG	42.27	32.31	36.63
(SIGARRA	PER	57.39	62.14	59.67
+	PLC	37.35	51.38	43.26
SecHAREM)	TME	71.33	74.91	73.08
	VAL	80.19	82.52	81.34

Pirovani, J. and Oliveira, E. Studying the Adaptation of Portuguese NER for Different Textual Genres. The Journal of Supercomputing, 2021.

# APLICAÇÕES

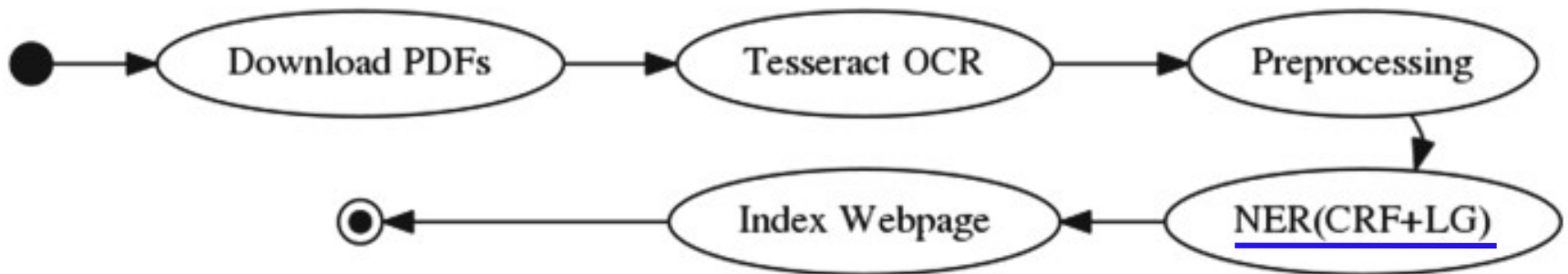
## ■ Geração de Perguntas e Respostas

<b>Sentença</b>	A Revolução Francesa trouxe significativos avanços no tratamento da questão da tortura.
<b>Sentença após <u>REN</u></b>	A <EM ID="Cap14-297" <u>CATEG="ACONTECIMENTO"</u> > <u>Revolução Francesa</u> </EM> trouxe significativos avanços no tratamento da questão da tortura.
<b>Sentença após análise sintática</b>	NP[A Revolução Francesa] VP[trouxe] NP[significativos avanços] PP[em o] NP[tratamento] PP[de a] NP[questão] PP[de a] NP[tortura]
<b>Análise sintática com REN</b>	NP[A ACONTECIMENTO:Revolução Francesa] VP[trouxe] NP[significativos avanços] PP[em o] NP[tratamento] PP[de a] NP[questão] PP[de a] NP[tortura]
<b>Regra usada na GQ</b>	<NP <ACONTECIMENTO>> <VP> ... → Qual acontecimento <VP> ...?
<b>Questão gerada</b>	Qual acontecimento trouxe significativos avanços no tratamento da questão da tortura?

Pirovani, J., Spalenza, M., & Oliveira, E. (2017, October). Geração automática de questões a partir do reconhecimento de entidades nomeadas em textos didáticos. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE) (Vol. 28, No. 1, p. 1147).

# APLICAÇÕES

## ■ Indexação Semântica para Pesquisa mais Eficiente de Documentos



Pirovani, J. P., Nogueira, M., & de Oliveira, E. (2018, September). Indexing Names of Persons in a Large Dataset of a Newspaper. In International Conference on Computational Processing of the Portuguese Language (pp. 147-155). Springer, Cham.



# APLICAÇÕES

## ■ Extração de Relacionamentos

---

<code>is_father(abraham,isaac).</code>	
<code>is_father(isaac,jacob).</code>	<code>is_father(jacob,judah).</code>
<code>is_father(judah,perez).</code>	<code>is_mother(tamar,perez).</code>
<code>is_father(judah,zerah).</code>	<code>is_mother(tamar,zerah).</code>
<code>is_father(perez,hezrom).</code>	<code>is_father(hezrom,ram).</code>
<code>is_father(ram,amminadab).</code>	<code>is_father(amminadab,nahshon).</code>
...	
<code>is_father(abiud,eliakim).</code>	<code>is_father(eliakim,azor).</code>
<code>is_father(azor,zadok).</code>	<code>is_father(zadok,akim).</code>
<code>is_father(akim,elihud).</code>	<code>is_father(elihud,eleazar).</code>
<code>is_father(eleazar,matthan).</code>	<code>is_father(matthan,jacob).</code>
<code>is_father(jacob,joseph).</code>	<code>is_father(joseph,jesus).</code>
<code>is_father(abraham,david).</code>	<code>is_father(david,jesus_christ).</code>
<code>is_mother(mary,jesus).</code>	

---

---

Oliveira, E., Dias, G., Lima, J., & Pirovani, J. P. C. (2021, October). Using Named Entities for Recognizing Family Relationships. In Anais do IX Symposium on Knowledge Discovery, Mining and Learning (pp. 24-32). SBC.

---

# CONCLUSÕES

- Ferramentas como CRF+LG podem ser usadas com o objetivo de extrair metainformação de documentos automaticamente.
  - A abordagem usada no CRF+LG também pode ser usada para extrair outros tipos de ENs e estruturas nos documentos.
-

---

# REFERÊNCIAS

- AMARAL, D. O. F. d. O Reconhecimento de Entidades Nomeadas por Meio de Conditional Random Fields para a Língua Portuguesa. Dissertação (Mestrado) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil, 2013.
  - Santos, C.N.d., Guimaraes, V.: Boosting named entity recognition with neural character embeddings. In: Proceedings of the Fifth Named Entities Workshop, ACL 2015. pp. 25–33 (2015).
  - Paumier, S. Unitex 3.3 User Manual, 2021. Disponível em: <https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf>. Acesso em: 22/10/2021.
  - Pirovani, J. and Oliveira, E. (2021). Studying the Adaptation of Portuguese NER for Different Textual Genres. The Journal of Supercomputing, pages 1–17.
-

---

# Obrigada!

Dúvidas?  
[juliana.campos@ufes.br](mailto:juliana.campos@ufes.br)

---