

# **Segundo HAREM Workshop**

## PROPOR 2008: International Conference on Computational Processing of Portuguese Language

# **Sistema SeRELeP para o reconhecimento de relações**

Mírian Bruckschen  
[mirian.bruckschen@gmail.com](mailto:mirian.bruckschen@gmail.com)

Renata Vieira  
[renata.vieira@pucrs.br](mailto:renata.vieira@pucrs.br)

José Guilherme Camargo de Souza  
[joseguilhermecs@gmail.com](mailto:joseguilhermecs@gmail.com)

Aveiro, setembro de 2008

# Roteiro

- Introdução
- Sistema SeRELeP
  - Visão geral
  - Reconhecimento de relações
  - Resultados e discussão
- Discussão sobre a trilha ReRelEM
- Considerações finais

Com o crescimento e evolução do número de esforços no sentido de reconhecer relações entre entidades, recentemente, havia pouca atenção ao reconhecimento de relações entre entidades em português. Além disso, o reconhecimento de relações entre entidades é subjetivo, dado que não existe uma definição universal.

Neste contexto, surgem os sistemas SeRELeP e ReRelEM, que são sistemas de Reconhecimento de Entidades e Relações entre Entidades, respectivamente. O sistema SeRELeP tem como objetivo bastante específico: reconhecer relações entre entidades mencionadas, com nome próprio ou não, em textos em português.

Na tentativa de ampliar o sistema SeRELeP para o reconhecimento de relações entre entidades em português, o sistema ReRelEM (Reconhecimento de Entidades e Relações entre Entidades) foi criado. O sistema ReRelEM é um sistema de Processamento de Linguagem Natural (PLN) que realiza a avaliação, referente ao reconhecimento de relações entre entidades, de textos em português.

No que refere-se à identificação de entidades (de referência e de correferência), este sistema é capaz de atender a uma variedade de tipos de entidades, com aplicabilidade nas implicações da construção automática de relações entre entidades.

Neste trabalho, é apresentado o sistema ReRelEM. O sistema ReRelEM recebe como entrada textos em português, processa-os com o sistema SeRELeP e faz inferências linguísticas. Já existem outras discussões sobre o reconhecimento de relações entre entidades neste trabalho, que não são abordadas aqui.

O restante do documento aborda os conceitos básicos e trabalhos relacionados ao reconhecimento de relações entre entidades, o sistema projetado e desenvolvido, os resultados preliminares; e a Seção Conclusão e Trabalhos Futuros.

# Introdução

- Oportunidade de participação no Segundo HAREM<sup>1</sup> (HAREM, 2007)
    - Trilha de reconhecimento de relações entre EMs<sup>2</sup> e experiência do grupo sobre correferência

<sup>1</sup> HAREM é uma Avaliação de Reconhecedores de Entidades Mencionadas

## <sup>2</sup> Entidades Mencionadas

# Visão geral (1/3)

- Sistema SeRELeP<sup>6</sup>
  - Desenvolvido em Python
  - Objetivo: reconhecer relações entre EMs previamente identificadas pelo PALAVRAS
  - Entrada: arquivos do *corpus* pré-processados pelo analisador PALAVRAS e o conversor Tiger2XCES
    - SeRELeP Tools: pacote associado que realiza tarefas de conversão entre ferramentas
  - Saída: arquivos anotados com EMs e suas relações

<sup>6</sup> Sistema de Reconhecimento de RELações entre EMs da Língua Portuguesa

# Visão geral (2/3)

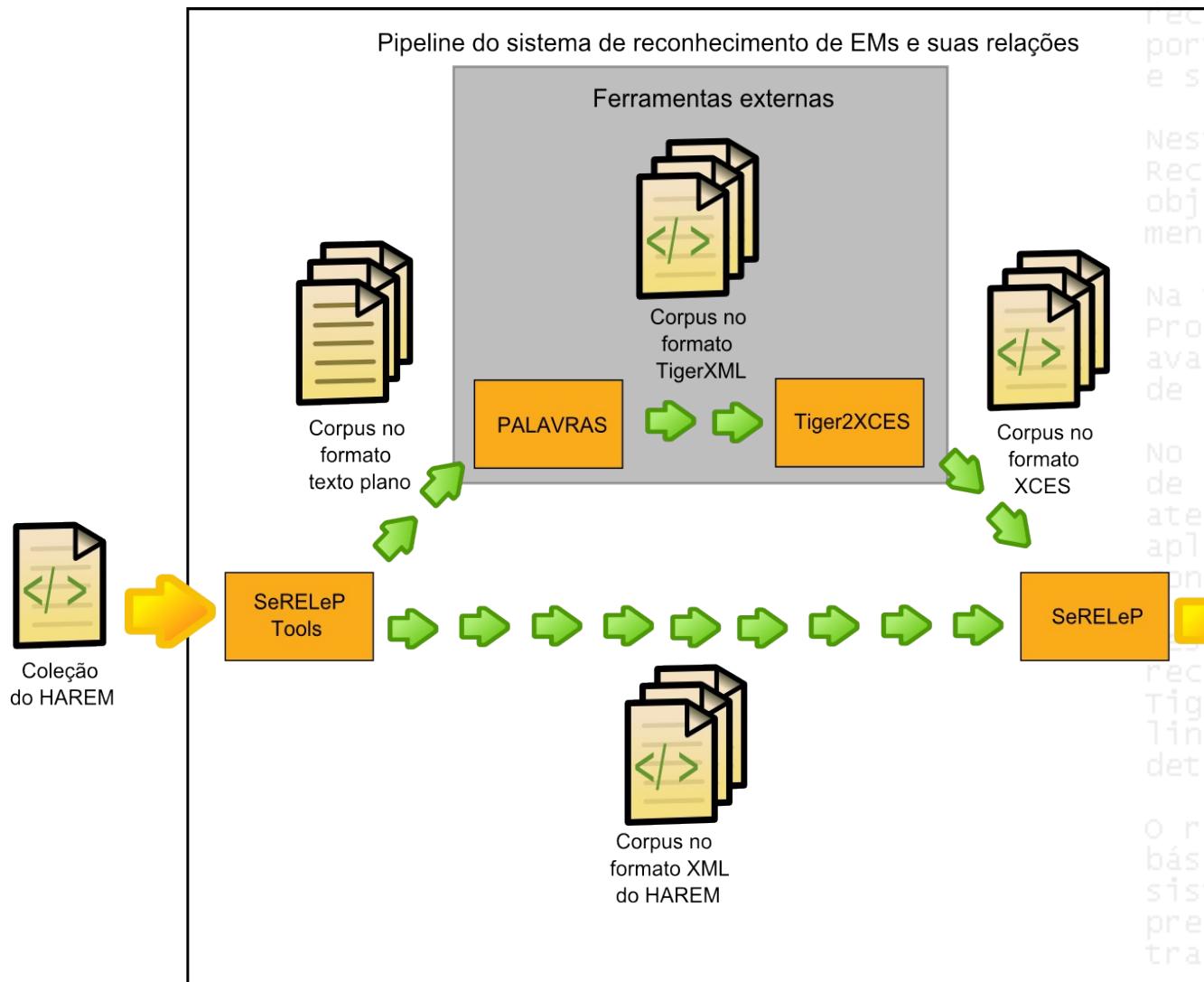


Figura 1. Processo de anotação automática de EMs e relações da coleção do HAREM  
Sistema SeRELeP

## Visão geral (3/3)

<DOC DOCID="cha-6282">

De Lisboa para Cascais e de novo para Lisboa, a mulher do Presidente da República Francesa, Danielle Mitterrand, passou ontem um dia ocupado na divulgação da versão portuguesa do «Passaporte Europeu contra o Racismo», um documento pessoal em que cada um se compromete simbolicamente a «resistir a qualquer acto de racismo». Recebida de manhã por Maria Barroso, que deu o seu patrocínio a esta iniciativa da Civitas, associação

<DOC DOCID="cha-6282">

De **Lisboa** para Cascais e de novo para **Lisboa**, a mulher do **Presidente da República Francesa**, **Danielle Mitterrand**, passou ontem um dia ocupado na divulgação da versão portuguesa do «**Passaporte Europeu contra o Racismo**», um documento pessoal em que cada um se compromete simbolicamente a «resistir a qualquer acto de racismo». Recebida de manhã por **Maria Barroso**, que deu o seu patrocínio a esta iniciativa da **Civitas**,

Figura 2. Trechos de entrada e saída do *pipeline*

# Reconhecimento de relações (1/3)

- O processo de reconhecimento de relações baseia-se na informação fornecida nos arquivos XCES
  - Se é EM (prop) ou não (PALAVRAS)
  - Etiquetação semântica (PALAVRAS)

# Reconhecimento de relações (2/3)

- Existe uma relação da etiqueta semântica atribuída à classificação da EM no HAREM
  - Com base nessa etiquetação, são definidas as heurísticas para reconhecimento das relações
  - Exemplo: “hum” ou “groupofficial” no PALAVRAS: PESSOA no HAREM

# Reconhecimento de relações (3/3)

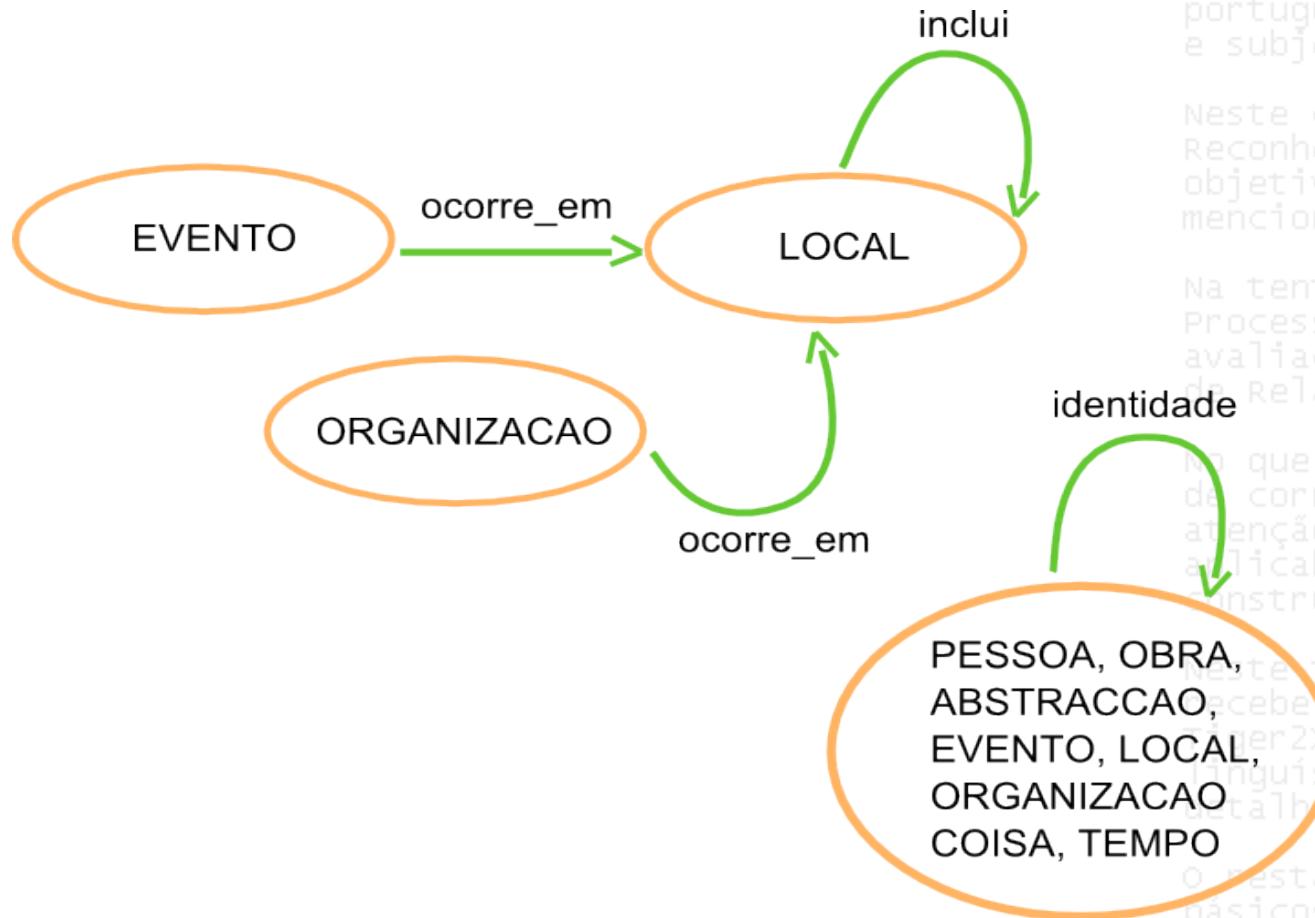


Figura 3. Relações e classes semânticas das EMs no SeRELeP.

# Reconhecimento das relações

- ident
  - *exact match*, sigla, parte do nome em caso de PESSOA (Silva e José da Silva)
- inclui
  - tratada entre entidades de LOCAL, procura entidades da mesma sentença que estejam na forma <entidade1> (...) em (...) <entidade2>
- ocorre\_em
  - pedaço do nome do evento ou organização é nome de local
    - São Leopoldo Fest **ocorre em** São Leopoldo
  - entidade de local na mesma frase que evento ou organização
  - local mais próximo no texto

# Resultados e discussão (1/5)

	Precisão	Abrangência	Medida F
<b>Identificação</b>	82%	60%	69%

Tabela 1. HAREM clássico (PALAVRAS)

# Resultados e discussão (2/5)

	<b>Precisão</b>	<b>Abrangência</b>	<b>Medida F</b>
<b>Aval. Relações</b>	58%	31%	40%

Tabela 2. ReReIEM (SeRELeP) sem a relação “outra”

# Resultados e discussão (3/5)

- identidade
  - 89% de **precisão**, 55% de **abrangência**, 68% de **medida F**
- inclui
  - 54% de **precisão**, 11% de **abrangência**, 18% de **medida F**
    - Poucas regras utilizadas, a baixa abrangência era esperada
- ocorre\_em (localização)
  - 36% de **precisão**, 27% de **abrangência** e 31% de **medida F**
    - Um problema com relação à precisão: muitos casos marcados pelo sistema como **ocorre\_em** eram **outra**, na verdade

# Resultados e discussão (4/5)

- Abordagem simples
- Melhorias planejadas
  - ident
    - Nomes “similares” (*edit distance*)
    - Utilização de informação de aposto
    - Tradutores *online*
    - Wikipedia
      - Artigo na Wikipedia pt: Lisboa
      - Artigo na Wikipedia en: Lisbon
  - inclui e ocorre\_em
    - Consulta a bases de dados (ontologias, Wikipedia, *gazetteers*)
      - Relações de difícil tratamento somente com regras linguísticas

# Resultados e discussão (5/5)

- Diferenças entre identificação e classificação
  - Vagueza, talvez?
  - A relação presente entre EMs vagas é somente **outra**
  - SeRELeP não contempla vagueza
    - “Brasil” é sempre LOCAL
    - “Brasil” sempre **inclui** “Porto Alegre” (mesmo que este se refira à seleção brasileira de futebol...)

com o crescimento e evolução do número de esforços no sentido de recentemente, havia poucos trabalhos em português. Além disso, o termo é subjetivo, dado que não é sempre o mesmo.

Neste contexto, surgem questões: Reconhecedores de Entidades? Um objetivo bastante específico, que não é mencionado, com nome próprio.

Uma tentativa de ampliar o campo de aplicação para o Processamento de Linguagem Natural (PLN) é a avaliação, referente ao desempenho do sistema, de Relações entre Entidades.

No que refere-se à identificação de entidades (de referência), este trabalho não faz a atenção de diversos esforços que já foram feitos, mas é importante para a construção automática de sistemas de NLP.

Neste trabalho, é apresentado o sistema que recebe como entrada texto em português e faz inferências linguísticas. Já existem outros sistemas que fazem o mesmo, mas o detalhe neste trabalho, é que o sistema é feito de forma mais simples.

O restante do documento aborda os resultados obtidos, os sistemas básicos e trabalhos relacionados, o sistema projetado e desenhado, os resultados preliminares; e a Seção Conclusão e Trabalhos Futuros.

# Discussão sobre a trilha ReReEM

- Resolução de correferência
  - Relação **ident**
    - Exemplo de cadeia de correferência #1

- **Felix\_Mirabel** , pesquisador que liderou o grupo
- **Mirabel**
- o pesquisador
- **ele**

Com o crescimento e evolução do número de esforços no sistema, recentemente, havia pouca documentação em português. Além disso, o sistema é subjetivo, dado que não

Na tentativa de ampliar o Processamento de Língua, a avaliação, referente ao de Relações entre Enunciado e

Neste trabalho, é apresentado o que é recebido como entrada de texto por um sistema de inferência linguística. Já existem trabalhos que detalham este processo.

# Discussão sobre a trilha ReReLEM

- Resolução de correferência
- Relação **ident**
  - Exemplo de cadeia de correferência #2

- pesquisadores de a Universidade de Wisconsin-Madison (EUA)
- A equipe liderada por Yoshihiro\_Kawaoka
- Os cientistas
- o grupo de Kawaoka

Com o crescimento e evolução do número de esforços no setor, recentemente, havia pouca literatura em português. Além disso, o termo é subjetivo, dado que não é formalizado.

Neste contexto, surgem as Relações Reconhecedores de Entidades (RE), com objetivo bastante específico, mencionadas, com nome próprio.

Na tentativa de ampliar o campo de aplicação do Processamento de Linguagem Natural, a avaliação, referente ao uso das Relações entre Entidades.

No que refere-se à identificação de correferência (que é o caso de identidade de referentes), este trabalho tem atenção de diversos aspectos, como a aplicabilidade nas implicações da construção automática de referentes.

Neste trabalho, é apresentado o sistema que recebe como entrada texto em português e faz inferências linguísticas. Já existem detalhes sobre o sistema em detalhe neste trabalho, que é o que se segue.

O restante do documento descreve os aspectos básicos e trabalhos relacionados ao sistema, seu projeto e desenvolvimento preliminares; e a Seção 5, que descreve o trabalho.

# Discussão sobre a trilha ReReLEM

- Resolução de correferência
- Relação **ident**
  - Exemplo de cadeia de correferência #3

- Brasileiros
- Os brasileiros – Gilson Rambelli, Paulo Bava de Camargo e Flávio Rizzi, pesquisadores de o Museu de Arqueologia e Etnologia (MAE) de a USP –
- Eles
- os arqueólogos
- os três brasileiros
- os arqueólogos

# Discussão sobre a trilha ReReLEM

## Outras relações

- subconjunto/parte-de

- Exemplo

- a Via Láctea
      - o Sol
      - a Terra

- outra

- Exemplo

- a estrela
      - o que sobra

Com o crescimento e evolução do número de esforços no sistema, recentemente, havia pouca documentação em português. Além disso, o sistema é subjetivo, dado que não é possível garantir a exatidão.

Neste contexto, surgem as Relações entre Entidades Reconhecedores de Entidades. O objetivo é obter uma relação entre entidades mencionadas, com nome próprio.

Na tentativa de ampliar o sistema de Processamento de Linguagem Natural, a avaliação, referente ao sistema de Relações entre Entidades, é realizada.

No que refere-se à identificação de cor referência (referência), este sistema atende a diversos espe- cíficos, com aplicabilidade nas impor- tantes tarefas de construção automática de sistemas de linguagem natural.

Neste trabalho, é apresentado o sistema que recebe como entrada textos em português e faz inferências linguísticas.

Tiger2XCES e faz inferências linguísticas. Já existem detalhes sobre o sistema de inferências e detalhe neste trabalho, que é apresentado.

O restante do documento é dividido em seções básicos e trabalhos relacionados ao sistema, o sistema projetado e desenhado, os resultados preliminares; e a Seção de conclusões.

# Corpus Summ-it

- Summ-it v3.0 – já disponível para *download*
  - 50 textos jornalísticos de ciências retirados da Folha de São Paulo
  - Composição
    - textos originais e tarjados (com informações relevantes do texto)
    - arquivos XML com anotação morfossintática (PALAVRAS)
    - **arquivos com anotação manual de correferência (MMAX)**
    - arquivos XML com anotação RST (RSTTool)
    - sumários automáticos e manuais
  - <<http://www.inf.pucrs.br/~linatural/procacosa.htm>>

# Análise de resultados

- Resultados preliminares são razoáveis
  - Mas podem (e devem!) ser melhorados
- Propostas de aplicação do sistema SeRELeP
  - Geração automática de ontologias de determinado tipo de EMs e relações a partir de conjuntos de textos
  - **ident** pode auxiliar nas tarefas de resolução de correferência e enriquecimento de sumários automáticos
  - SeRELeP-Olympics
    - Geração de *hot topics* (com o auxílio da relação **ident**) para um portal de notícias sobre as Olimpíadas
      - “Cielo” e “César Cielo” se referem à mesma entidade, portanto devem aumentar o *ranking* desta entre os *hot topics*

# Trabalhos futuros

- Melhorias nos algoritmos de reconhecimento de relações
  - Consulta a bases de dados externas, principalmente
- Ampliação da tarefa para incluir mais que EMs, objetivando cadeias de relações mais completas e informativas, focadas no tipo de entidade

Com o crescimento e evolução do número de esforços no setor, recentemente, havia poucos trabalhos em português. Além disso, o resultado era subjetivo, dado que não havia uma base de dados para validação. Neste contexto, surgem os primeiros sistemas de Reconhecedores de Entidades (REs) com objetivo bastante específico, que foram mencionadas, com nome próprio, em artigos de revistas científicas.

Além da necessidade de ampliar o escopo das tarefas de Reconhecimento e Processamento de Linguagem Natural (RPLN) para a realização de inferências entre Entidades (REs), que se refere à identificação e classificação (REC), este trabalho também atende a diversos esforços de pesquisa que visam a aplicabilidade nas implicações entre Entidades. A aplicabilidade das implicações entre Entidades é uma questão que não é abordada de forma detalhada neste trabalho, mas é uma questão importante para a aplicabilidade das implicações entre Entidades.

Neste trabalho, é apresentado um sistema de Reconhecimento e Processamento de Linguagem Natural (RPLN) que recebe como entrada texto em português e faz inferências entre Entidades. Já existem sistemas de Reconhecimento e Processamento de Linguagem Natural (RPLN) que realizam inferências entre Entidades, mas não é abordado de forma detalhada neste trabalho, mas é uma questão importante para a aplicabilidade das implicações entre Entidades.

O restante do documento aborda os principais resultados obtidos, os sistemas projetados e desenvolvidos, as implicações entre Entidades e a Seção de Conclusão. O restante do documento aborda os principais resultados obtidos, os sistemas projetados e desenvolvidos, as implicações entre Entidades e a Seção de Conclusão.

# Agradecimentos

- Agradecemos imensamente à organização do Segundo HAREM pela oportunidade de participação, pela atenção, paciência e pela receptividade a todos nossos comentários e sugestões

# Referências

- Collovini, S.; Carbonel, T.; Fuchs, J. T.; Coelho, J. C.; Rino, L.; Vieira, R. (2007) *Summ-it: Um corpus com informações discursivas visando à summarização automática*. In: Anais do XXVII Congresso da SBC (TIL – V Workshop em Tecnologia da Informação e Linguagem Humana).
- HAREM. (2007) *HAREM: Reconhecimento de entidades mencionadas em português*. Disponível em: [http://acdc.linguateca.pt/aval\\_conjunta/HAREM/](http://acdc.linguateca.pt/aval_conjunta/HAREM/). Acesso em: junho de 2008.
- Norvig, P. (2008) *How to Write a Spelling Corrector*. Disponível em: <http://norvig.com/spell-correct.html>. Acesso em: agosto de 2008.
- Santos, D.; Cardoso, N. (eds.) (2007) *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, Portugal.
- Souza, J. G. C. (2007) *Resolução automática de correferência aplicada à língua portuguesa*. Monografia (Graduação). Curso de Ciência da Computação, UNISINOS. Brasil.

# **Segundo HAREM Workshop**

## PROPOR 2008: International Conference on Computational Processing of Portuguese Language

# **Sistema SeRELeP para o reconhecimento de relações**

Mírian Bruckschen  
[mirian.bruckschen@gmail.com](mailto:mirian.bruckschen@gmail.com)

Renata Vieira  
[renata.vieira@pucrs.br](mailto:renata.vieira@pucrs.br)

José Guilherme Camargo de Souza  
[joseguilhermecs@gmail.com](mailto:joseguilhermecs@gmail.com)

Aveiro, setembro de 2008