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

Renata Vieira
renata.vieira@pucrs.br

José Guilherme Camargo de Souza
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

Introdução

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

[1] HAREM é uma Avaliação de Reconhecedores de Entidades Mencionadas
[2] Entidades Mencionadas

Visão geral (1/3)

- Sistema SeRELeP[6]
  - 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

[6] Sistema de Reconhecimento de RELações entre EMs da Língua Portuguesa

Visão geral (2/3)

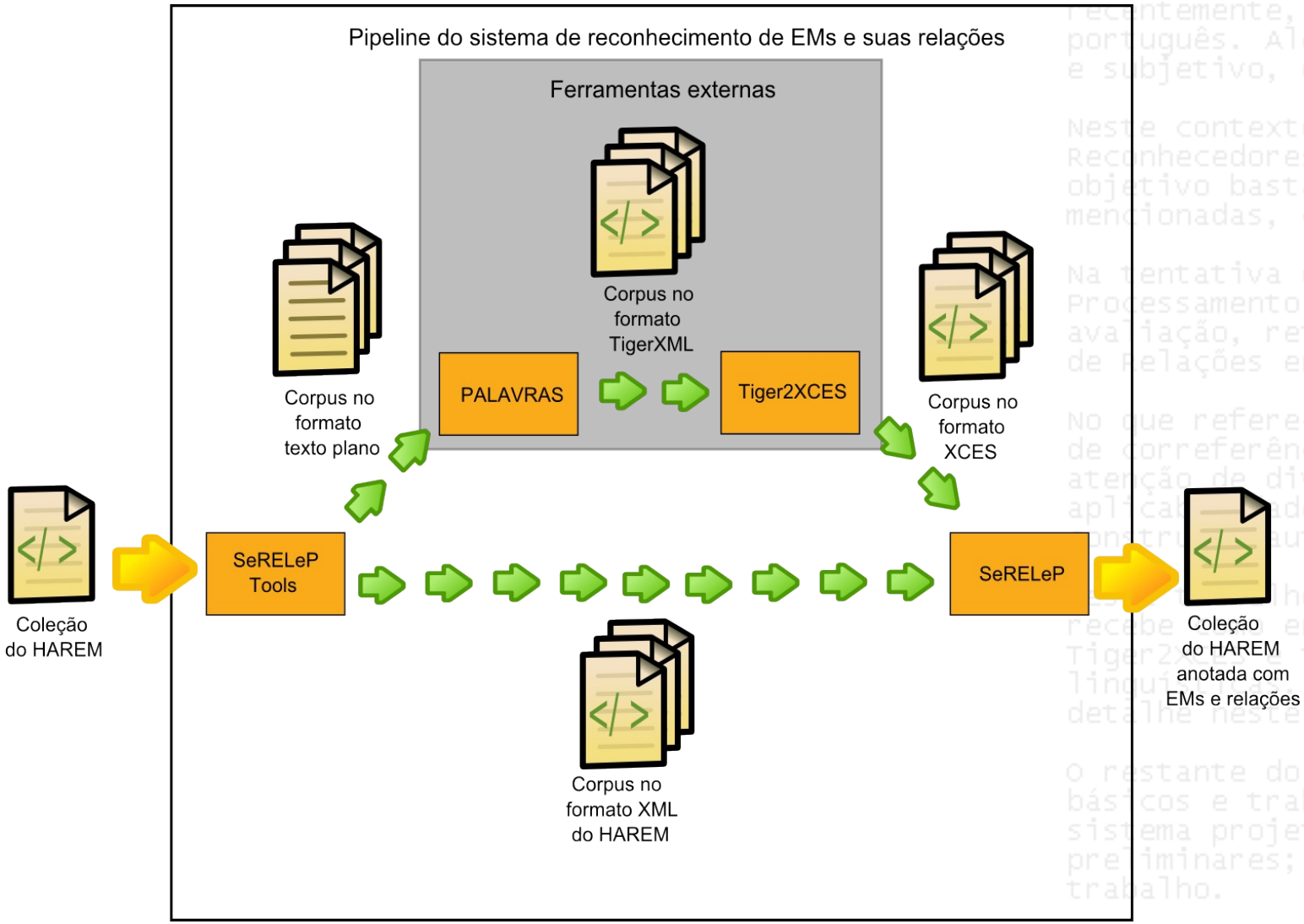

Figura 1. Processo de anotação automática de EMs e relações da coleção do HAREM

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 <EM ID="cha-6282-EM_1">Lisboa</EM> para Cascais e de novo
para <EM ID="cha-6282-EM_2" COREL="cha-6282-EM_1" TIPOREL=
"ident">Lisboa</EM>, a mulher do <EM ID="cha-6282-EM_3">
Presidente da República Francesa</EM>, <EM ID="cha-6282-EM_4"
>Danielle Mitterrand</EM>, passou ontem um dia ocupado na
divulgação da versão portuguesa do «<EM ID="cha-6282-EM_5">
Passaporte Europeu contra o Racismo</EM>», um documento
pessoal em que cada um se compromete simbolicamente a
«resistir a qualquer acto de racismo». Recebida de manhã por
<EM ID="cha-6282-EM_6">Maria Barroso</EM>, que deu o seu
patrocínio a esta iniciativa da <EM ID="cha-6282-EM_7" COREL=
"cha-6282-EM_12" TIPOREL="ocorre_em">Civitas</EM>,
```

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)

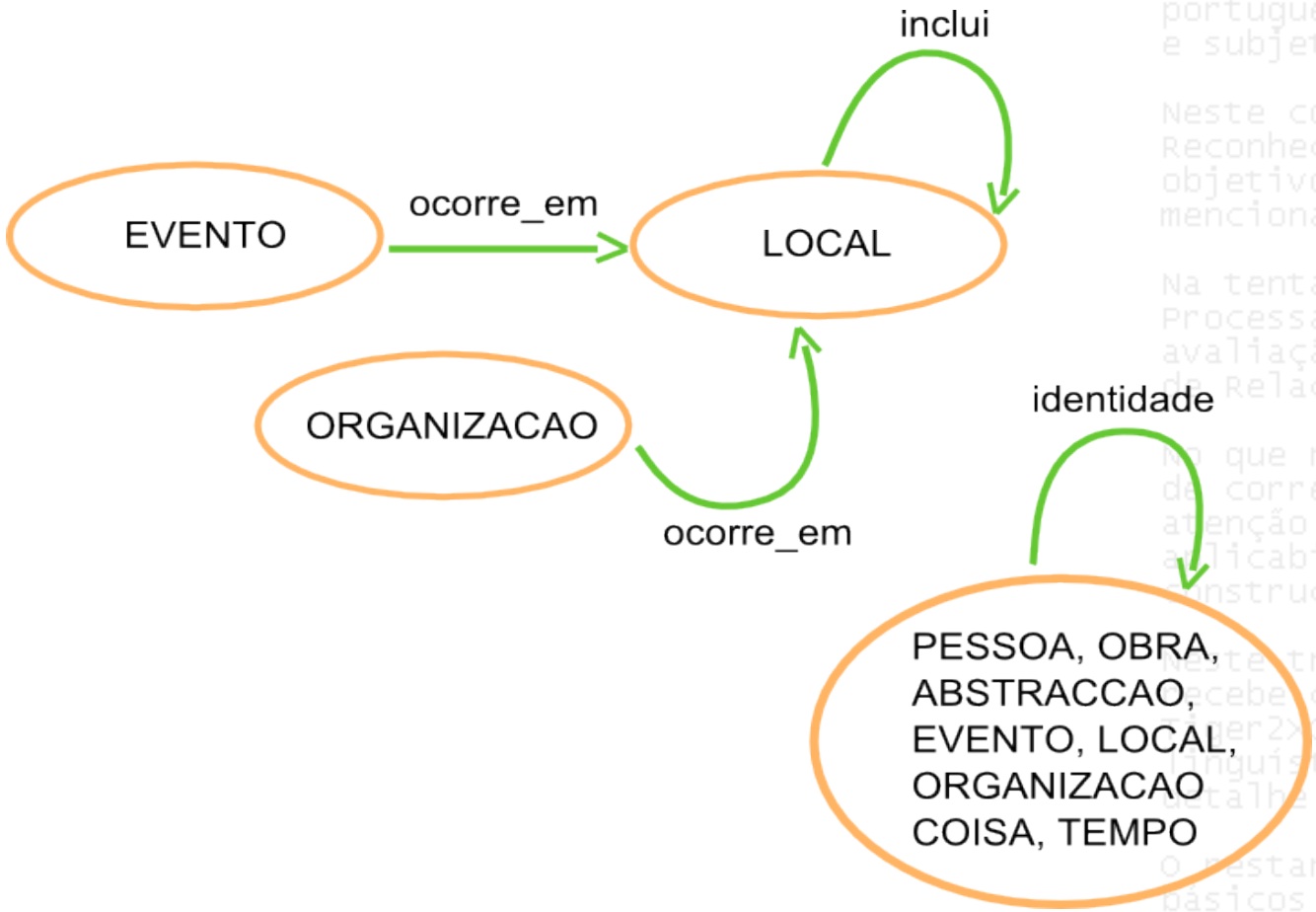

Figura 3. Relações e classes semânticas das EMs no SeRELeP

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
Identificação	82%	60%	69%

Tabela 1. HAREM clássico (PALAVRAS)

Resultados e discussão (2/5)

	Precisão	Abrangência	Medida F
Identificação	68%	35%	46%
Classificação	57%	30%	39%

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

Resultados e discussão (3/5)

- ident
  - 87% de **precisão**, 54% de **abrangência**, 66% de **medida F**
- inclui
  - 56% de **precisão**, 11% de **abrangência**, 19% de **medida F**
    - Poucas regras utilizadas, a baixa abrangência era esperada
- ocorre_em (localização)
  - 28% de **precisão**, 21% de **abrangência** e 24% 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...)

Discussão sobre a trilha ReRelEM

- Resolução de correferência
- Relação **ident**
  - Exemplo de cadeia de correferência #1
    - Felix_Mirabel , pesquisador que liderou o grupo
    - Mirabel
    - o pesquisador
    - ele

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

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

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

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 sumarização 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/. 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 lingua 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

Renata Vieira
renata.vieira@pucrs.br

José Guilherme Camargo de Souza
joseguilhermecs@gmail.com

Aveiro, setembro de 2008