Capítulo 14

Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas

Mírian Bruckschen, José Guilherme Camargo de Souza, Renata Vieira e Sandro Rigo

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 14, p. 247–260.

Neste capítulo, é apresentado um sistema focado no reconhecimento de relações entre EM. As subtarefas de identificação e classificação das EM são realizadas pelo analisador sintático PALAVRAS (Bick, 2000), deixando como tarefa do sistema aqui apresentado somente o reconhecimento das relações entre estas EM.

O sistema faz inferência das relações a partir de regras heurísticas simples, que consideram apenas informações presentes no próprio texto e informações adicionais providas pelo PALAVRAS. Assim, o sistema não faz uso de bases de conhecimento adicionais para o reconhecimento das relações, resolvendo a tarefa através de regras linguísticas e de posicionamento das EM em cada texto analisado.

O restante do documento está organizado da seguinte forma: a seção 14.1 relata brevemente a experiência anterior do grupo em análise de correferência, que foi parte da motivação para participação do Segundo HAREM; a seção 14.2 faz uma descrição detalhada do sistema projetado e desenvolvido, assim como uma discussão dos resultados; e a seção 14.3 finaliza o documento com considerações finais e futuras direções do trabalho.

14.1 Trabalhos relacionados e motivação

Uma das motivações que nos levou a participar do processo do Segundo HAREM foi a experiência anterior do nosso grupo na tarefa de resolução de correferência – que pode ser facilmente associada à relação de identidade proposta pelo Segundo HAREM.

Apesar de a proposta do HAREM ser uma tarefa diferenciada da nossa experiência com resolução de correferência, tratando exclusivamente nomes próprios mas também outras relações além de identidade, reconhecemos a importância da avaliação conjunta, que nunca ocorreu no contexto de resolução de correferência para a língua portuguesa antes do Segundo HAREM.

A abordagem para resolução de correferência, tal como adotada em Souza et al. (2008), leva em consideração não apenas as entidades mencionadas mas todos os tipos de sintagmas nominais referenciais presentes em um texto: indefinidos, definidos, pronomes e nomes próprios, mas sem dar enfoque à distinção de categorias (pessoa, local, organização). Assim, dado um conjunto de sintagmas nominais de um texto, o sistema tem por objetivo agrupar os sintagmas em cadeias que evocam a mesma entidade. O processo de resolução de correferência desenvolvido é formado por três momentos: (i) geração de pares de sintagmas nominais, (ii) classificação dos pares quanto a sua anaforicidade e (iii) agrupamento dos pares anafóricos em cadeias.

Junto com a geração dos pares de sintagmas, são verificadas características que são consideradas no processo de classificação automática por aprendizado. Essas características são informações morfossintáticas, posicionais e semânticas. O classificador é induzido por aprendizado de máquina supervisionado com base em um corpo anotado, o Summ-it[1] (Collovini et al., 2007).

O classificador indica quais pares de sintagmas são relacionados por anaforicidade. A partir dos pares identificados, os conjuntos de sintagmas correferentes são formados.

Existem outras abordagens bastante conhecidas e bem-sucedidas, mas geralmente aplicadas a outras línguas (o inglês, em particular) (Soon et al., 2001).

A participação no HAREM foi uma experiência complementar e inspiradora. Acreditamos que uma revisão e união das duas abordagens seja produtiva e importante para o

[1] Disponível em: http://www.inf.pucrs.br/~linatural/procacosa.htm.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<DOC DOCID="2ght33"> (...)
A relação do tecno com o binômio homem-máquina, no diálogo com <EM ID="2ght33-EM_2">Jeff Mills
    </EM>, o legendário produtor de tecno de <EM ID="2ght33-EM_3">Detroit</EM> nos leva a
    impressionantes insights do impacto que a terceira onda tem causado na paisagem
    contemporânea. <EM ID="2ght33-EM_4" COREL="2ght33-EM_3" TIPOREL="ident">Detroit</EM>,
    essa "cidade portátil", virtualizada na minimalista batida de um sequenciador automático,
     profetiza em sua música - que já nos deu a <EM ID="2ght33-EM_5">Motown</EM>, <EM ID="2
    ght33-EM_6">Stooges</EM>, e <EM ID="2ght33-EM_7">MC5</EM> - o zeitgeist deste início de
    milênio. (...)
</DOC>
```

Figura 14.1: Trecho de arquivo de saída do SeRELeP

entendimento do problema e aperfeiçoamento da tarefa. Focar nos diferentes tipos de categorias de cadeias, por exemplo, pode ser uma maneira de organizar os sistemas, com isso melhorar os resultados e refinar e sua avaliação.

14.2 SeRELeP: Sistema de reconhecimento de RElações em textos de Língua Portuguesa

SeRELeP é um **S**istema de reconhecimento de **RE**lações em textos de **L**íngua **P**ortuguesa. Foi desenvolvido visando a participação na pista de reconhecimento de relações entre EM (ReRelEM) do Segundo HAREM (consulte-se o capítulo 4 para uma apresentação da pista). A ferramenta, sua metodologia de desenvolvimento e resultados são detalhados nesta seção.

14.2.1 Visão geral

O SeRELeP propõe-se a demarcar as relações de identidade (`ident`), ocorrência (`ocorre_em`) e inclusão (`inclui`) entre EM conforme as diretrizes do Segundo HAREM, que se encontram no apêndice C. Partindo do reconhecimento e classificação de EM efetuados pelo analisador PALAVRAS, o sistema processa a coleção de textos do HAREM e retorna a mesma coleção com a anotação das relações entre EM.

O sistema tem como entrada o arquivo de texto da coleção do HAREM (em formato XML[2]) e seus respectivos arquivos em formato XCES[3]. Para obtenção dos arquivos neste formato, é necessário o processamento do corpo anotado em TigerXML (König e Lezius, 2003) pelo conversor Tiger2XCES (Bruckschen et al., 2008b). Como saída, o SeRELeP devolve um arquivo com o texto já marcado com as EM e suas relações, também em formato XML. A figura 14.1 traz um trecho de um arquivo de saída como exemplo. Na figura 14.2 é ilustrado todo o processo de anotação automática.

Nessa figura, SeRELeP é o sistema identificador de relações entre as EM, e o SeRELeP Tools é um conjunto de pequenos programas auxiliares, necessários à etapa de pré-processamento. A entrada do processo é um arquivo XML no formato do HAREM, fornecido no início da participação da avaliação conjunta. Este arquivo contém diversos textos individuais.

[2] *eXtensible Markup Language*
[3] XML CES: *Corpus Encoding Standard for XML*, conforme http://www.xces.org/

Os textos são extraídos pelo SeRELeP Tools em dois formatos: texto plano, que é a entrada para o PALAVRAS, e XML do HAREM, que é entrada para o SeRELeP efetivamente. Além do XML do HAREM, o SeRELeP ainda precisa de outra entrada, que são os textos anotados pelo PALAVRAS no formato XCES.

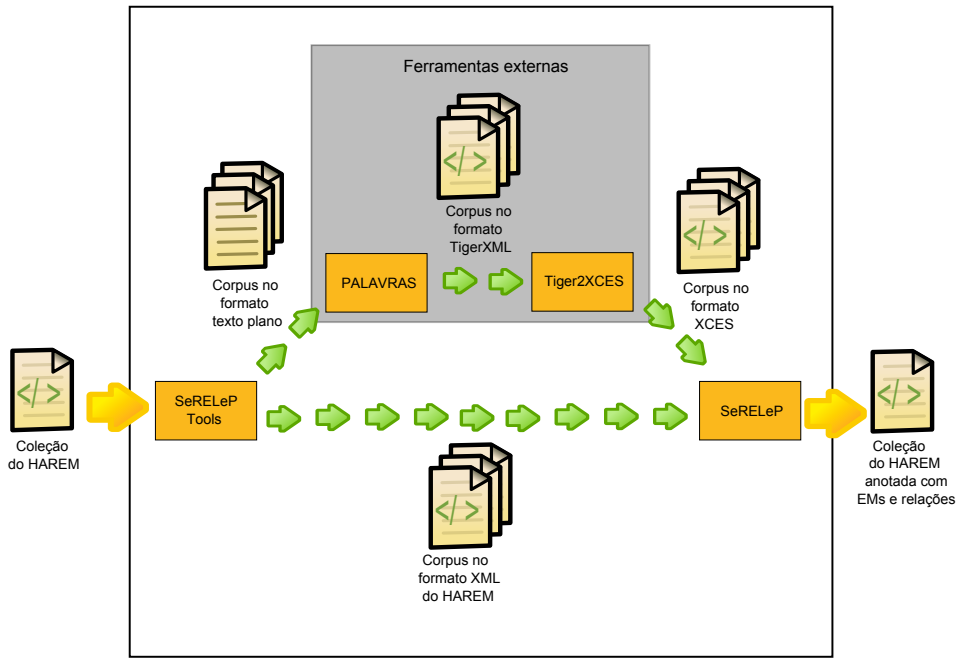

Figura 14.2: Processo de anotação automática de EM e relações da coleção do HAREM

A utilização do formato XCES é devida à diminuição de complexidade de interpretação por analisadores automáticos, pois é um formato proposto a fim de atender a vários critérios, levantados como necessários a um formato padrão de representação de informação linguística. Alguns destes critérios são expressividade, independência de mídia, adequação semântica, simplicidade (e legibilidade por humanos, tanto quanto possível), incrementabilidade e extensibilidade (Ide e Romary, 2004). O formato adotado baseia-se nessas diretivas e é apresentado em detalhe no relatório do projeto PLN-BR (Bruckschen et al., 2008a). Seguindo este formato, a anotação do PALAVRAS é codificada em três arquivos XCES a partir de cada texto original: `token`, `pos` e `phrase`. Cada um destes arquivos representa um nível de anotação linguística. Os itens de informação são delimitados pelo elemento `struct`, e todas as suas características por elementos `feat`. A estrutura do documento anotado é predominantemente vertical ao invés de horizontal; existem muitos elementos no arquivo, mas cada um destes possui poucos atributos. Esta característica favorece a criação de analisadores para este tipo de arquivo, e de igual forma torna a informação mais clara para leitura por seres humanos (Bruckschen et al., 2008a).

O arquivo `token` identifica as unidades lexicais (ou átomos). A cada unidade corresponde um elemento XML ao qual é dado um identificador, e nos campos `from` e `to` são informados o início e o fim deste elemento no texto, de acordo com sua posição em carac-

```
<?xml version="1.0" standalone="yes" ?>
<cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
 <struct to="5" type="token" from="0">
  <feat name="id" value="t1" />
  <feat name="base" value="Mills" />
 </struct>
 <struct to="7" type="token" from="6">
  <feat name="id" value="t2" />
  <feat name="base" value="é" />
 </struct>
 <struct to="10" type="token" from="8">
  <feat name="id" value="t3" />
  <feat name="base" value="um" />
 </struct>
 <struct to="16" type="token" from="11">
  <feat name="id" value="t4" />
  <feat name="base" value="homem" />
 </struct>
 <struct to="22" type="token" from="17">
  <feat name="id" value="t5" />
  <feat name="base" value="calmo" />
 </struct>
 <struct to="23" type="token" from="22">
  <feat name="id" value="t6" />
  <feat name="base" value="." />
 </struct>
</cesAna>
```

Figura 14.3: Trecho de arquivo de *token* no formato XCES

teres. No exemplo 14.1, o átomo *um* é identificado e vai da posição 9 à 10.

(14.1) Mills é um homem calmo

O arquivo `POS` (*part-of-speech*) representa a informação de nível morfossintático (as etiquetas semânticas também são representadas neste arquivo). Os elementos de `token` são referenciados em cada elemento de `POS`.

Finalmente, o arquivo `phrase` descreve a informação de nível sintático: sentenças e sintagmas, identificação de sujeitos, predicados e objetos. Os grupos sintagmáticos são identificados como intervalos de elementos `token`.

A seguir, são ilustrados os trechos de arquivos de `token`, `POS` e `phrase` do formato XCES referenciados. Estes são as figuras 14.3, 14.4 e 14.5, respectivamente. Todos eles ilustram a sentença 14.1.

O SeRELeP e seu módulo de programas auxiliares SeRELeP Tools foram desenvolvidos em Python[4], utilizando a biblioteca SAX[5] para processamento de arquivos XML.

A definição das classes para representação de informação linguística baseou-se no conversor Tiger2XCES, por sua vez, escrito na linguagem de programação Java.

O arquivo de programa principal é o `SeRELeP/serelep.py`, que processa os arquivos XCES (`token`, `pos` e `phrase`) e executa os dois principais métodos do processo de reconhecimento automático de relações entre as EM: i) a procura por nomes próprios e sua classificação, e ii) a inferência das relações a partir de critérios especificados nos métodos

[4] http://python.org/

[5] *Simple API for XML*, conforme disponível em http://www.python.org/doc/lib/module-xml.sax.html.

```
<?xml version="1.0" standalone="yes" ?>
<cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
 <struct type="pos">
  <feat name="id" value="pos1" />
  <feat name="class" value="prop" />
  <feat name="tokenref" value="t1" />
  <feat name="canon" value="Mills" />
  <feat name="gender" value="M" />
  <feat name="number" value="S" />
  <feat name="complement" value="hum" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos2" />
  <feat name="class" value="v-fin" />
  <feat name="tokenref" value="t2" />
  <feat name="canon" value="ser" />
  <feat name="complement" value="fmc" />
  <feat name="complement" value="mv" />
  <feat name="tense" value="PR" />
  <feat name="person" value="3S" />
  <feat name="n_form" value="VFIN" />
  <feat name="mode" value="IND" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos3" />
  <feat name="class" value="art" />
  <feat name="tokenref" value="t3" />
  <feat name="canon" value="um" />
  <feat name="gender" value="M" />
  <feat name="number" value="S" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos4" />
  <feat name="class" value="n" />
  <feat name="tokenref" value="t4" />
  <feat name="canon" value="homem" />
  <feat name="gender" value="M" />
  <feat name="number" value="S" />
  <feat name="semantic" value="Hattr" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos5" />
  <feat name="class" value="adj" />
  <feat name="tokenref" value="t5" />
  <feat name="canon" value="calmo" />
  <feat name="gender" value="M" />
  <feat name="number" value="S" />
  <feat name="complement" value="np-close" />
 </struct>
 <struct type="pos">
  <feat name="id" value="pos6" />
  <feat name="class" value="pu" />
  <feat name="tokenref" value="t6" />
 </struct>
</cesAna>
```

Figura 14.4: Trecho de arquivo de POS no formato XCES

```
<?xml version="1.0" standalone="yes" ?>
<cesAna version="1.0.4" xmlns="http://www.xces.org/schema/2003">
 <struct to="t5" type="phrase" from="t1">
  <feat name="id" value="phr1" />
  <feat name="cat" value="s" />
  <feat name="function" value="" />
 </struct>
 <struct to="t5" type="phrase" from="t1">
  <feat name="id" value="phr2" />
  <feat name="cat" value="fcl" />
  <feat name="function" value="STA" />
 </struct>
 <struct to="t1" type="phrase" from="t1">
  <feat name="id" value="phr3" />
  <feat name="cat" value="prop" />
  <feat name="function" value="S" />
 </struct>
 <struct to="t2" type="phrase" from="t2">
  <feat name="id" value="phr4" />
  <feat name="cat" value="v-fin" />
  <feat name="function" value="P" />
 </struct>
 <struct to="t5" type="phrase" from="t3">
  <feat name="id" value="phr5" />
  <feat name="cat" value="np" />
  <feat name="function" value="Cs" />
  <feat name="head" value="t4" />
 </struct>
</cesAna>
```

Figura 14.5: Trecho de arquivo de *phrase* no formato XCES

apropriados. O processo de reconhecimento de relações é descrito em detalhe na subseção 14.2.2.

Além das classes de representação da informação linguística do texto e analisadores para cada um dos tipos de entrada (`token`, `POS` e `phrase` em XCES, e XML do HAREM), foram desenvolvidos, de forma individual, os métodos para cada uma das técnicas para reconhecimento de relações (descritos na subseção 14.2.2). A figura 14.6 representa as relações entre principais arquivos de código-fonte e métodos do sistema.

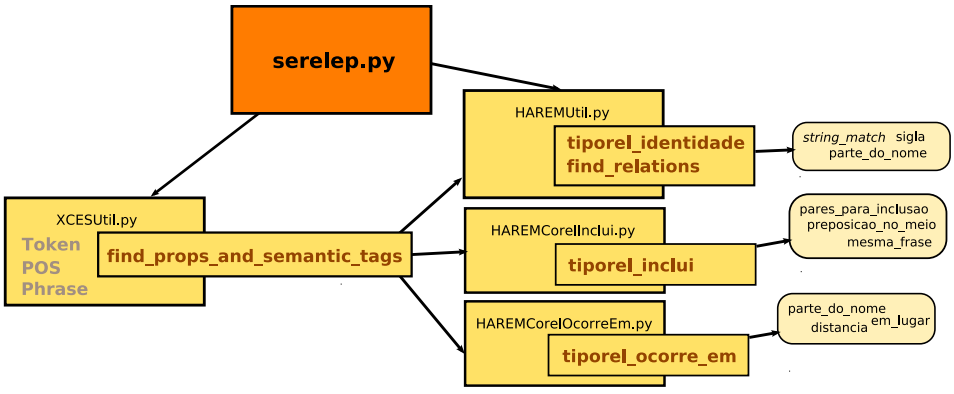

Figura 14.6: Visão geral do código-fonte e métodos principais do SeRELeP

Todos os métodos auxiliares usam primariamente dicionários e listas, que são estruturas básicas do Python, e bastante otimizadas. As informações utilizadas (como as etiquetas do PALAVRAS e sua correspondência com as categorias do HAREM) são também definidas na forma de dicionários. Esta decisão por certo influenciou positivamente no desempenho computacional do sistema, cuja única etapa mais demorada é o pré-processamento.

Além da utilização de estruturas básicas da linguagem, é preciso considerar o fato de que as regras utilizadas são simples e facilmente otimizáveis. Depois do pré-processamento, o processo todo não ultrapassa 10 minutos para o reconhecimento de relações e anotação de toda a coleção do Segundo HAREM – 1048 textos (tempo verificado num portátil Core Duo 1.6 GHz, 1GB DDR, executando a distribuição de Ubuntu GNU/Linux 7.10, com núcleo (em inglês, *kernel*) de Linux 2.6).

Com relação ao sistema, convém observar que foi desenvolvido de forma modularizada para facilitar a inclusão de novas técnicas e regras como, por exemplo, o uso de outras técnicas de reconhecimento de relações através da utilização de bases de dados externas.

14.2.2 Reconhecimento de relações entre entidades mencionadas

A marcação `prop` (nome próprio) do analisador PALAVRAS é utilizada para identificação e delimitação das EM no texto, e as suas etiquetas semânticas usadas para a classificação.

As etiquetas (e sua correspondência com as categorias do HAREM) foram as utilizadas na primeira edição do HAREM (Bick, 2007). A figura 14.7 ilustra estas etiquetas e sua correspondência com as categorias. O significado de cada etiqueta está disponível na página de documentação do PALAVRAS-VISL[6].

As associações realizadas entre etiquetas do PALAVRAS e classes do HAREM são diretas, não houve um tratamento adicional para a vagueza, a etiqueta `civ` sempre será referente a `LOCAL`, por exemplo.

Figura 14.7: Etiquetas semânticas do PALAVRAS e as categorias do HAREM

Conforme já comentado, o HAREM propõe quatro relações entre EM: `ident` (identidade), `inclui`, `ocorre_em` e `outra`. Destas, o SeRELeP trata as três primeiras. O tratamento de cada uma destas relações atualmente é detalhada mais adiante nesta seção.

[6] http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html

Convém notar que as heurísticas descritas abaixo fazem uso da etiquetação semântica do pré-processamento no que refere-se à categorização das EM. A figura 14.8 expressa na forma de um grafo dirigido as relações `ident`, `inclui` e `ocorre_em` entre as EM pertencentes a estas classes conforme tratado pelo SeRELeP. Salientamos que o grafo representado é uma simplificação, e que o nó `ORGANIZACAO/LOCAL` são de facto dois e `PESSOA/.../TEMPO` representa cinco nós distintos cada um deles referente a um tipo de EM (de outro modo, seria possível existir uma relação `ident`, por exemplo, entre `ORGANIZACAO` e `LOCAL`, o que como veremos mais adiante não acontece).

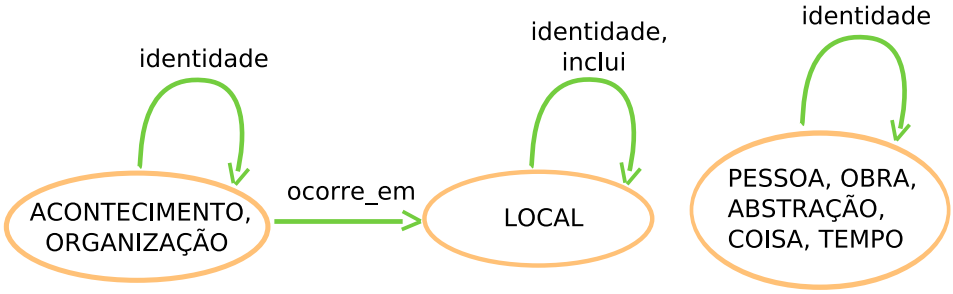

Figura 14.8: Relações e classes semânticas das EM conforme tratado pelo SeRELeP

As relações são reconhecidas numa determinada ordem, começando a partir da relação `ident`, já que as demais dependem dessa. Depois desta, são avaliadas as relações `inclui` e `ocorre_em`.

A relação `ident` é atribuída a EM que se referem a uma mesma entidade no mundo. A atribuição desta relação dá-se através das seguintes regras:

i) comparação direta de cadeias de carateres, isto é, se as EM possuem exatamente o mesmo nome;

ii) se uma é sigla da outra, isto é, se uma das EM retoma as iniciais da outra e possui mais de um caractere (a EM *DA* poderia ser relacionada com *Departamento de Artes*, por exemplo, mas *A*, isoladamente, não seria relacionada a *Artes*);

iii) se as EM comparadas forem da classe `PESSOA` e parte do sintagma de uma for igual ao sintagma da outra (como *Carmem* e *Carmem Miranda*, por exemplo). Além disso, as EM devem pertencer à mesma categoria semântica (uma EM de `ACONTECIMENTO` só pode estabelecer relação de identidade com outra EM de `ACONTECIMENTO`, por exemplo).

O exemplo 14.2 traz um trecho de texto com referências a uma mesma entidade que devem ser marcadas como possuindo a relação identidade entre si: *São Leopoldo* e *SL*.

(14.2) *São Leopoldo* é uma cidade localizada na região metropolitana de Porto Alegre, no Rio Grande do Sul. Dentre as diversas atrações da cidade, localiza-se em SL o Museu do Trem, um museu ferroviário, e um teatro recém inaugurado pela prefeitura junto à biblioteca municipal.

Já a relação `inclui`, tratada entre EM de `LOCAL` e simétrica da relação `incluído`, é estabelecida mediante as seguintes regras:

i) as duas EM não podem ter relação `ident` entre si;

ii) devem estar na mesma sentença;

iii) deve haver uma preposição que denote inclusão, como *em*, *no* e *na*.

Pode-se observar que são regras simples. Desta forma, não foi uma surpresa a baixa abrangência dos resultados desta relação. Os únicos casos de relações que seriam encontrados seriam os realmente explícitos, onde numa mesma sentença fosse referenciada a entidade incluída e a que a inclui.

O exemplo 14.3 traz um trecho de texto com entidades que estão ligadas pela relação de inclusão. No exemplo temos três entidades, e três relações de inclusão (uma delas implícita): *Rio Grande do Sul* `inclui` *São Leopoldo*, *Brasil* `inclui` *Rio Grande do Sul* e, consequentemente, *Brasil* `inclui` *São Leopoldo*.

(14.3) *São Leopoldo* é localizado no Rio Grande do Sul, no **Brasil**.

Finalmente, a relação `ocorre_em` é tratada entre EM de `ACONTECIMENTO` e `LOCAL` ou de `ORGANIZACAO` e `LOCAL`. As regras obedecidas por ela são verificadas na seguinte ordem:

i) se houver uma EM de `LOCAL`, cujo sintagma seja parte do sintagma da EM de `ACONTECIMENTO` ou `ORGANIZACAO` verificada, essa EM inserida é relacionada à EM de `LOCAL` em questão (como em *Brigada Militar de Porto Alegre* `ocorre_em` *Porto Alegre*);

ii) se isso não acontecer, é verificada a existência de uma EM de `LOCAL` na mesma sentença da EM de `ACONTECIMENTO/ORGANIZACAO` analisada. Se existir, esta EM de `ACONTECIMENTO/ORGANIZACAO` será relacionada a esta EM de `LOCAL` através da relação `ocorre_em`;

iii) se não, busca a EM de `LOCAL` mais próxima dentro do texto (se houver) para relacionar com a EM de `ACONTECIMENTO/ORGANIZACAO` analisada.

Com este conjunto de heurísticas, o SeRELeP obteve, entre os outros sistemas, o melhor resultado no reconhecimento da relação `ocorre_em`.

São ilustrados dois casos de ocorrência (de EM de `ACONTECIMENTO`) no exemplo 14.4.

(14.4) Ocorre em *São Leopoldo* a São Leopoldo Fest, festa que se dá em homenagem à chegada dos imigrantes alemães fundadores da cidade, e que reúne participantes de todo o estado. Além disso, a cidade é palco anualmente da sua Feira do Livro, com a participação de escritores, seus fãs, e diversas personalidade do mundo literário.

14.2.3 Resultados

Na coleção final anotada pelo SeRELeP, não classificamos as EM explicitamente, somente as suas relações, e por este motivo os resultados da classificação de EM (HAREM clássico) não são aqui exibidos. O principal motivo para não ser feita esta anotação foi o foco na tarefa de reconhecimento de relações. Sendo esta classificação resultado do processamento de uma ferramenta externa à desenvolvida e aqui descrita, não incluimos essa informação

na coleção anotada devolvida para avaliação – apesar de ter sido usada na inferência das relações.

Ainda assim, mostramos na tabela 14.1 os resultados da tarefa de identificação de EM no cenário total com avaliação estrita de ALT obtidos pela corrida SeRELeP_1. Esses são resultados obtidos pelo PALAVRAS, que mantém-se entre os melhores sistemas que participaram do Segundo HAREM na tarefa de identificação de entidades.

Tabela 14.1: Resultados oficiais do HAREM clássico (PALAVRAS)

	Precisão	**Abrangência**	**Medida F**
Identificação	0,82	0,60	0,69

Tabela 14.2: Comparativo dos resultados oficiais da pista do ReRelEM

	Precisão	**Abrangência**	**Medida F**
REMBRANDT_1	0,58	0,44	0,50
SeRELeP_1	0,58	0,31	0,40
SeRELeP_no	0,57	0,30	0,39
REMBRANDT_2	0,27	0,48	0,35
REMBRANDT_3_corr	0,25	0,48	0,32

Tabela 14.3: Comparativo dos resultados oficiais da pista do ReRelEM por relação

		Precisão	**Abrangência**	**Medida F**
REMBRANDT	ident	0,77	0,69	0,73
	inclui	0,32	0,33	0,33
	ocorre_em	0,40	0,13	0,20
SEI-Geo	ident	-	-	-
	inclui	0,92	0,30	0,45
	ocorre_em	-	-	-
SeRELeP	ident	0,89	0,55	0,68
	inclui	0,54	0,11	0,18
	ocorre_em	0,36	0,27	0,31

Os resultados da tarefa de reconhecimento de relações são ilustrados na tabela 14.2, que traz uma comparação da avaliação do reconhecimento de relações, no cenário seletivo do ReRelEM que inclui todas as relações menos outra, das cinco corridas melhores posicionadas. Diferentes corridas significam diferentes anotações, potencialmente por sistemas com ajustes e parâmetros diferentes. No caso particular do SeRELeP a corrida SeRELeP_1 distingue-se da corrida SeRELeP_no, não oficial, pela correção de um erro de delimitação das entidades identificado e reportado pela organização. A tabela 14.3 compara os resultados do SeRELeP com os dos restantes sistemas nos cenários seletivos do ReRelEM constituídos por cada uma das relações tratadas.

A relação com melhores resultados do SeRELeP é claramente a ident. Atribui-se este desempenho ao fato de que regras simples, como as utilizadas e descritas neste documento, já abrangem boa parte das relações entre EM. Ainda assim, o melhor classificado

nesta relação é o sistema REMBRANDT, enquanto o SeRELeP lidera no reconhecimento da relação `ocorre_em` e o SEI-Geo, a relação `inclui`.

Quanto ao reconhecimento de relações pelo SeRELeP, algumas questões ainda devem ser tratadas em maior detalhe, como apelidos não relacionados ao nome original (um exemplo seria *Pequena Notável* e *Carmem Miranda*, que têm relação de identidade não-detectada pelo sistema) e o mesmo nome com pequenas diferenças de grafia, comumente causados por erros de digitação (*Maria de Costa* e *Maria da Costa*).

As relações `inclui` e `ocorre_em` possuem resultados inferiores à relação `ident`.

O reconhecimento das relações, sobretudo nesses casos, pode ter sido afetado por problemas de classificação das EM[7]. Um exemplo disso é a marcação de EM de `LOCAL` tais como *Biblioteca Victor Civita* como `ORGANIZACAO`, no texto cujo trecho é ilustrado no exemplo 14.5. Em 14.6, por sua vez, a EM *Broadway* não é indicada na coleção dourada como um lugar, mas sim um grupo de pessoas, pertencendo à categoria semântica `PESSOA`, e não `LOCAL`. Já de acordo com o analisador sintático, a classificação resultante é `LOCAL`.

(14.5) Ele, que fez consultoria dos textos presentes na mostra, vai realizar uma palestra gratuita na sexta-feira, às 19h30, sobre a vida e a obra da artista na *Biblioteca Victor Civita*, no Memorial.

(14.6) *Carmen Miranda* conquistou a Broadway

Por outro lado, algumas EM corretamente classificadas tiveram relações identificadas incorretamente. No exemplo 14.7, temos algumas EM marcadas, e as seguintes relações identificadas: *África do Sul* `inclui` *Durban* (correta) e *Durban* `inclui` *Portugal* (incorreta). É possível perceber facilmente que o filtro simples aplicado a esta sentença (entidade de `LOCAL`, seguida pela preposição *em* e por outra entidade de `LOCAL`) ocasionou o problema, muito embora as entidades tivessem sido corretamente classificadas.

(14.7) Sua mãe casa-se pela segunda vez em 1895 por procuração, na *Igreja de São Mamede* em Lisboa, com o Comandante João Miguel Rosa cônsul de *Portugal* em Durban (África do Sul), o qual havia conhecido um ano antes.

Este é um dos exemplos que ilustra o motivo pelo qual entendemos que a utilização de informação externa ao texto (como bases de dados e ontologias de domínio – neste caso, geográfico) seriam acréscimos interessantes ao sistema apresentado.

14.3 Considerações finais

Motivadas por um fator crítico bastante rígido (tempo), as principais técnicas para o reconhecimento de relações foram baseadas em heurísticas simples, resultantes da análise dos exemplos no corpo.

Entendemos nossa participação como uma experiência modesta, uma vez que utilizamos ferramentas já existentes para o REM. A escolha do formato XCES deve-se à experiência na participação no projeto PLN-BR.

[7] Apesar de não ser anotada na coleção a classificação das EM, esta informação foi utilizada para a inferência das relações.

No entanto, o desenvolvimento do sistema demandou um tempo e esforço razoáveis, e conseguimos enviar a coleção do Segundo HAREM para participar da avaliação conjunta somente nos últimos dias do prazo final.

Tivemos uma surpresa muito positiva ao receber os resultados no final de agosto de 2008, e ver que o SeRELeP era bastante competitivo (com precisão melhor na maioria dos casos) com os outros dois sistemas concorrrentes na pista do ReRelEM. Ambos os sistemas, comparativamente, eram trabalhos mais maduros.

Atribuímos os bons resultados, em parte, à tarefa de REM muito bem-executada pelo PALAVRAS. Apesar dos problemas percebidos por nós no pré-processamento – e nossa queixa do quanto isso influenciou os nossos resultados, considerando que muitos dos erros provêm dessa etapa – também temos que mencionar que o PALAVRAS ainda é o melhor analisador morfossintático para a língua portuguesa.

Ainda, é importante lembrar que algumas técnicas para o reconhecimento das relações eram mais difíceis, e implementadas até em caráter experimental, mas que a maioria delas foi criada tendo com base a análise subjetiva de textos, procurando por padrões. As regras lingüísticas para reconhecimento da relação de identidade, por outro lado, ficaram óbvias tão logo nos debruçamos sobre as diretrizes da tarefa e vimos o quanto cada relação deveria abranger.

Com certeza, apesar de já positivos, estes resultados poderiam ser bastante aprimorados. Como trabalho futuro, pretendemos utilizar outras informações morfossintáticas adicionais como aposto e predicativo, que podem auxiliar principalmente na relação `ident`. Além disso, pretende-se usar algoritmos de distância de edição, como o utilizado para a correção ortográfica, a fim de tratar os casos onde há pequenas diferenças de grafia nos nomes das entidades (Navarro, 2001).

Também pretendemos explorar bases de conhecimento externas tais como ontologias de domínio, corpos ou conteúdos disponíveis na rede (inicialmente a partir da Wikipedia[8]), tal como realizado pelo sistema REMBRANDT (ver capítulo 11). Acreditamos que isso deverá melhorar substancialmente os resultados das relações `inclui` e `ocorre_em`.

Outra frente de pesquisa é a utilização do SeRELeP no auxílio à geração de ontologias a partir de processamento de textos. Para isso, consideramos importante abranger outras entidades do texto na composição dos relacionamentos, tais como substantivos comuns, a exemplo da resolução de correferência clássica.

Gostaríamos ainda de realizar uma avaliação do SeRELeP desconsiderando informação de vagueza, a fim de verificar como seriam os resultados se a avaliação considerasse por exemplo a categoria país englobando seus aspectos de localidade, admistrativos e de população.

Pode-se observar aplicações bastante interessantes para sistemas de reconhecimento de relações. Grande parte destas aplicações pode ser relacionada à extração e classificação de informações, como a procura de notícias sobre alguma pessoa ou lugar, posição (em inglês, *ranking*) de entidades mais evidentes, sistemas de identificação de assuntos ou notícias relacionadas, sistemas de respostas automáticas baseados em consultas à rede e análise semântica dos resultados destas consultas.

Como extensão deste trabalho, está sendo desenvolvido o sistema SeRELeP-Olympics (Bruckschen et al., 2008c), que trata do uso do reconhecimento da relação de identidade, inicialmente, para a lista de tópicos mais frequentes (em inglês, *hot topics*) num portal de

[8] http://pt.wikipedia.org/

notícias sobre as Olimpíadas. Nesse sistema, todas as notícias que referenciem a mesma entidade (como *Cielo*, *Cesar Cielo* e *Cesar Cielo Filho*, que nomeiam o nadador brasileiro ganhador da medalha de ouro) são marcadas como relacionadas àquela entidade, independente da forma com que ela foi referenciada. E, nesse caso, são consideradas as referências feitas em diferentes documentos. No futuro, pretende-se incluir nessa aplicação outras relações, como a de inclusão (por exemplo, ocorrências de *Pequim* deveriam aumentar a posição de *China*).

Agradecimentos

Agradecemos ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e à FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul) pelo apoio no desenvolvimento deste trabalho.